Modelling the Effects of Speech Rate Variation for Automatic Speech Recognition

Der Technischen Fakultät der Universität Bielefeld

zur Erlangung des Grades einer

Doktor-Ingenieurin

vorgelegt von

Britta Wrede

Bielefeld - Juni 2002

Britta Wrede, M.A.
AG Angewandte Informatik
Technische Fakultät
Universität Bielefeld
email: bwrede@techfak.uni-bielefeld.de

Abdruck der genehmigten Dissertation zur Erlangung des akademischen Grades Doktor-Ingenieurin (Dr.-Ing.). Der Technischen Fakultät am 5.6.2002 vorgelegt von Britta Wrede.

Gutachter:

Dr. Gernot A. Fink
Prof. William Barry

Prüfungsausschuss:

Prof. Ipke Wachsmuth
Dr. habil. Gernot A. Fink
Prof. William Barry
Prof. Gerhard Sagerer
Dr. Katharina Rohlfing

Gedruckt auf alterungsbeständigem Papier ∞ ISO 9706

Acknowledgments

This work would not have been possible without the help and support of many people. While many PhD students are left alone with their thesis I feel extremely lucky in having had not only a very competent but also responsible supervisor with Dr. habil. Gernot Fink who was constantly available and more than willing to discuss problems, developments or new ideas when they arrised. Without his fundamental knowledge of the intricacies of ESMERALDA many solutions realised within this thesis would not have been possible. Also, I am very grateful to Prof. William Barry who agreed to review this thesis in a very tight schedule and whose comments were much appreciated.

Many thanks are due to Dr. Jacques Koreman for reading and re-reading parts of this thesis. His sharp eye spotted phonetic inconsistencies and hazardous syntactical constructions of English. I am more than grateful for his valuable comments and detailed questions which were a great help for giving this work its internal structure.

I would also like to thank the Applied Computer Science Group at Bielefeld which has always been a fun place to work. Special thanks go to (”Script God”) Christoph Schillo who not only proved to be an infinite source of awk- and shell-script knowledge and helpful little tools but who was also a great office mate and friend. Also, many thanks go to Christian Bauckhage for making the office such a plaesant and inspiring place to work - not only during the usual office hours.

I deeply appreciated the interdisciplinary atmosphere in the graduate school ”Task-oriented communication”. Not only did I receive much help with the planning of my thesis but I also learned how to discuss my research in an interdisciplinary framework.

Last but not least I would like to thank Stephan for being such a great coach and for his affection and emotional support whenever I needed it.

Contents

1. Introduction

In the last few years advanced human computer interaction has become more and more widespread. In order to build more intuitive interfaces, a focus was put on speech as the most natural medium for communication. However, prevailing speech recognition systems still fail to show a reliable performance on spontaneous speech. Current research therefore aims to increase the robustness of such systems by focusing on spontaneous speech.

Many features that are characteristic for spontaneous speech and therefore distinguish it from other kinds of speech such as read or dictation speech, cause severe problems for speech recognition systems. As compared to read speech spontaneous speech exhibits many hesitations and variations in the rate of speech as speakers have to plan their further speech while speaking. Also, slips of the tongue and repairs occur and the speaking style tends to become less clear. Thus, spontaneous speech shows a high degree of variation which makes it challenging to be adequately modelled by speech recognition systems.

One dimension that has been identified to cause a high degree of variation and accordingly recognition errors is the speech rate. In natural dialogues it is often the case that a speaker changes his or her speed of articulation substantially during an utterance. The modelling of speech rate has therefore received considerable attention. However, approaches that deal with rate variations tend to be isolated developments that have not yet been incorporated into general speech recognition systems. They require elaborate methods for a relatively small gain. Thus, current speech recognition systems are not yet able to capture variations of the speech rate very well.

A predominant paradigm in the modelling of speech rate is the training of rate dependent models which is achieved by a separation of the training data into discrete rate classes. On the one hand these approaches succeed in increasing the performance of the system on small corpora, but on the other hand they suffer from certain shortcomings related with the separation and thereby reduction of the training data which makes them ineffective for more substantial systems.

In general these approaches are missing detailed knowledge about the underlying effects of speech rate on the acoustic characteristics of the signal. Little is known about the actual effects that have to be modelled for faster or slower speech and how to capture those effects. In order to provide detailed insights into the effects that the acoustic correlates of speech rate

variations have on current speech recognition approaches more research is needed.

1.1. Aim of this Work

The aim of this work is to provide detailed information about the influence of rate variations and its acoustic correlates on speech recognition by systematically realising and evaluating solutions to several aspects of relevant problems of rate modelling.

Acoustic-phonetic analyses of the spectral effects of speech rate variations show that there is a systematic relationship between the duration of the phonetic segments and the speech signal. They indicate that acoustic characteristics can be predicted by the speech rate or the duration of the segments. This implies that approaches for rate modelling can be enhanced by incorporating knowledge about these relationships. In order to profit from this systematic relationship, more information about the nature of these effects has to be derived. Therefore, an acoustic analysis is carried out in this work in order to determine what kinds of effects can be found in the speech signal due to rate variations. The analysis provides detailed information on how acoustic cues of rate variation can be captured from the signal in order to model them.

This information is used to investigate the effectiveness of the different acoustic cues. Extensive speech recognition experiments are carried out in order to analyse in detail relevant aspects of speech rate modelling. These analyses are realised by making use of the prevailing technique of rate-dependent models.

It is shown that the effects that variations of the speech rate have on speech recognition systems are complex. Therefore, the current work elaborates these interactions and gives guidelines for a general framework which permits a detailed and robust modelling of speech rate. However, the realisation of such a substantial system lies beyond the scope of this work.

1.2. Outline

The structure of the investigation is reflected in the structure of this work. Since methods and techniques from two disciplines are combined the Chapter 2 gives an introduction into the fundamental aspects of acoustic-phonetic investigations and the principles of automatic speech recognition. While both disciplines focus on the same subject they are based on different methods. In order to transfer results from one discipline to the other a short comparison of the techniques is given at the end of the chapter.

Chapter 3 will then give a more detailed survey of the acoustic-phonetic implications

that speech rate variations have on the spectral and temporal characteristics of the signal. Thereby some measures of rate variation and their effects are presented and discussed. These measures are important for the further investigations.

In Chapter 4 several approaches to speech rate modelling in automatic speech recognition are presented. Again, the focus is put on the measures of speech rate which play an important role in those approaches. Furthermore, different compensation techniques are explained and discussed in their effectiveness.

The overviews of the acoustic effects of rate variation and their modelling in speech recognition are discussed in Chapter 5. In this chapter implications for speech rate modelling are drawn and precise questions for the further investigation are derived.

Many of the results reported in the literature about the acoustic influence of speech rate variation are based on small corpora consisting of read speech. In contrast, in this work acoustic-phonetic analyses are made on a large corpus of spontaneous speech in order to derive a detailed picture of the acoustic characteristics that occur with speech rate variation. The results of this analysis are presented in Chapter 6.

In Chapter 7 a series of speech recognition experiments is reported that was carried out in order to address the questions raised in Chapter 5. The results of further investigations on the acoustic models are presented in Chapter 7. It is shown how a modelling of the underlying spectral effects of rate variation can be achieved. Finally, a summary of the work is given in Chapter 8.

2. Introduction to Phonetics and Speech Recognition

Before the effects of speech rate variation are discussed from the two different standpoints of acoustic-phonetic research and speech recognition a survey of the fundamental methods and principles applied in both disciplines will be given in this chapter. However, a complete overview of both disciplines will not be attempted. Rather, an emphasis will be on those aspects that are important in the further course of this work. For a general introduction to phonetics refer to [CY90] while a more detailed survey of phonetic techniques and research is given in [HL97]. Readers interested in an introduction to automatic speech recognition are referred to [HAJ90]. More detailed information about the search algorithms for continuous speech recognition can be found in [NO99].

2.1. Phonetic Aspects

In this section some basic phonetic terms that play an important role in the process of this work will be explained. They refer to the articulation as well as to the acoustic signal that it produces. Since there is a close relationship between articulation and the acoustic characteristics of the speech signal it is helpful to consider the basic principles of articulation. Therefore, a short survey of articulation and its effects on the speech signal is given in this section. In order to describe the acoustic characteristics of speech signals some basic methods that are used for acoustic analysis in phonetic investigations are explained. However, some general terms and aspects concerning the notation are discussed first.

2.1.1. Notation

In order to analyse speech it proved useful to rely on the notions of phonemes and phones. A *phoneme* is a distinctive sound which serves to differentiate words. This means that phonemes are language specific. One method to determine the phonemes of a language is to build minimal pairs. That are word pairs which are differentiated by only one phone. For example, the words ”bet” and ”pet” are a minimal pair since they differ only in their first

phone. This means that the sounds ”b” and ”p” represent distinctive phonemes because they differentiate between the two words.

However, phonemes can be pronounced quite differently. For example, the German phoneme /r/ has differing pronunciations in some dialects. In standard German it is pronounced as a uvular voiced fricative or trill[1] while in some dialects it is rather realised as an alveolar trill. Such variants are called *allophones* of a phoneme. More generally, the term *phone* refers to actually produced instances of a phoneme while the term phoneme refers to an abstract entity which is defined by its function of discriminating words. In order to distinguish between a phonemic and a phonetic representation a convention is followed according to which phonemic transcriptions are written in slashes // while phonetic representations of actually produced sounds are written between brackets []. In the course of this work mostly phonetic transcriptions will be used since the focus of the investigations lies on the actual realisations.

In order to represent sounds it is necessary to provide an inventory of symbols that are related to the different sounds. The most common representation which is generally used for phonetic transcriptions is the *International Phonetic Alphabet* (IPA) as defined by the International Phonetic Association [Ass49]. Within this alphabet symbols are defined for about all phones which can be observed in the known languages. The alphabet provides a framework where the symbols are defined according to the articulatory features of the sounds they represent.

In this work the *Speech Assessment Methods Phonetic Alphabet* (*SAMPA*) will be used. The SAMPA is a machine-readable phonetic alphabet which basically consists of a mapping of symbols of the IPA onto ASCII codes. A list of the German SAMPA inventory as used for the phonotypic transcriptions of the lexicon for the speech recognition experiments and examples of transcriptions is given in Appendix A.

2.1.2. Articulation

In order to understand certain effects that rate variation has on the speech signal it is helpful to take a look at the process of articulation. The *articulatory system* consists basically of three components, the respiratory system which provides the air stream, the laryngeal system which consists of the glottis and the vocal folds for the sound production and finally the vocal tract. The vocal tract comprises roughly all parts between the glottis and the lips or nasal cavity where the air streams past during articulation. The shape of the vocal tract determines the acoustic characteristics of the speech signal. It is modulated by the articulators.

[1] A trill is a dynamic articulation produced by the vibration of an articulator as a consequence of the air stream passing by it. It is different from a fricative in that the frequency of the vibration is considerably lower.

By modulating the shape of the vocal tract certain frequencies in the resonance spectrum of the fundamental sound are emphasized or attenuated.

A sound is produced by air streaming through the glottis and exciting the vocal folds. By this excitation the fundamental sound is generated which is modulated by the supra-glottal vocal tract. The vibration of the vocal folds is controlled via adjustments in vocal fold adduction and tension and the stiffness of the folds. The tension of the vocal folds reflects the speech effort and affects the fundamental source spectrum. For voiceless sounds the vocal folds are open so that no fundamental tone is produced. The sound is then produced by obstructing the vocal tract with an articulator and releasing the obstruction which causes a plosion or a friction.

The *vocal tract* consists mainly of the oral cavity and the *articulators* which modulate the shape of the vocal tract in order to emphasise or attenuate certain frequencies. Modulations of the sound production can also be achieved by opening or closing the nasal cavity. However, for the sake of illustration the nasal cavity will be neglected in this description. The most important articulator is the tongue. Due to its size and the structure of its intrinsic and extrinsic muscles it is able to modulate the area function — the open area in the vocal tract through which the air can flow — in almost all places of the oral cavity. For *vowels* this area remains open during the whole articulation phase. In the production of vowels the most extreme positions of the tongue are taken for the vowels `[a:]` `[i:]` and `[u:]`. The `[i:]` is produced with a high, fronted tongue body. This means that the body of the tongue is shifted towards the front of the oral cavity and is raised so that it almost touches the palatum. In contrast, `[u:]` is produced with a high but retracted, or backed, tongue body. For `[a:]` the tongue is lowered which results in a wide area function in the oral cavity but a narrowing of the pharynx.

Contrary to this, for *consonants* the area function is constricted in at least one part of the vocal tract. This constriction can be a total occlusion of the air stream as for plosives or a very narrow constriction for fricatives where the air passes with a relatively high velocity and causes air turbulences. For example, for `[t]` the tongue tip forms a strict closure at the alveolar ridge while for `[s]` only a very narrow constriction is built through which air passes with a relatively high pressure. This causes turbulences in the air stream which are perceived as a friction noise.

In order to relate articulation to the acoustic characteristics of the speech signal the *Source Filter Model* [Fan70] provides a signal based model of the articulatory processes (cf. Figure 2.1). According to this model the complex sound wave of the speech signal can be interpreted as the result of two components of the articulatory system. For voiced sounds the basic component, the *source* of the sound, is the vibration of the vocal folds which produces the fundamental tone. The frequency of this tone is called the *fundamental frequency* (F0). All voiced sounds consist of a harmonic spectrum. This means that the sound not only con-

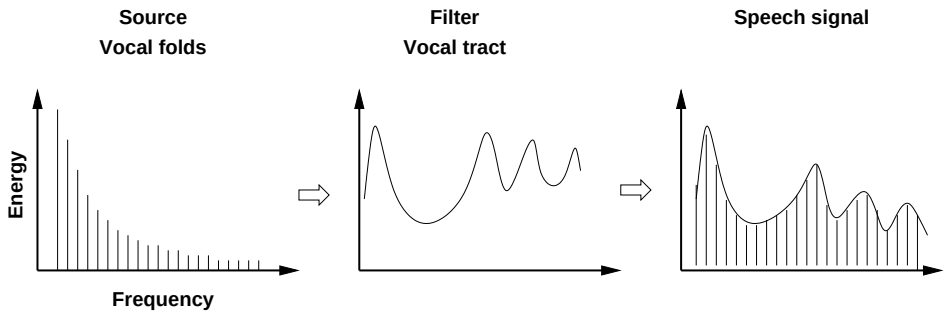

Figure 2.1.: Source Filter Model of speech production.

sists of a single frequency but of a whole spectrum where energy peaks occur at all integral multiples of the fundamental frequency with declining energy towards higher frequencies. These peaks are called *harmonics*. The second component of the source filter model represents the vocal tract which modulates the harmonic spectrum by emphasising or attenuating certain frequencies. This modification of certain frequency ranges is achieved by modifying the area function through which the sound wave travels in the vocal tract. By widening or narrowing the cross-sectional area in a certain place of the vocal tract a certain resonance frequency will be affected. Because of this function the vocal tract is often referred to as *filter*. The resulting signal can therefore be interpreted to represent the fundamental harmonic spectrum filtered by the vocal tract. Thus, on the one hand the spectrum of the speech signal conveys information about the sound source in terms of the fundamental frequency and the harmonic structure of the spectrum. On the other hand it contains information about the configuration of the vocal tract which is manifested in the frequency ranges that are emphasised or attenuated.

2.1.3. Acoustics

The source filter model indicates that the relevant characteristics of the speech signal can be found in its spectrum. The *spectrum* shows the energy with which different frequencies of the sound wave are represented in the speech signal. As can be seen in Figure 2.1 a spectrum is represented by the energy levels of the different frequency channels. The x-axis represents the frequency while the corresponding energy levels are indicated by the y-axis. In order to observe changes of the energy distribution over time *spectrograms* can be derived by combining the information of several consecutive spectra. In the resulting spectrogram the energy of the frequencies is plotted against time. Thus, while the x-axis represents time, the y-axis displays the frequencies. A high energy level of a certain frequency band at a certain point in time is represented by a dark point at the corresponding coordinate. Less energy is indicated by less intense colour. Figure 2.2 shows the spectrogram of the word ”Flugzeiten”

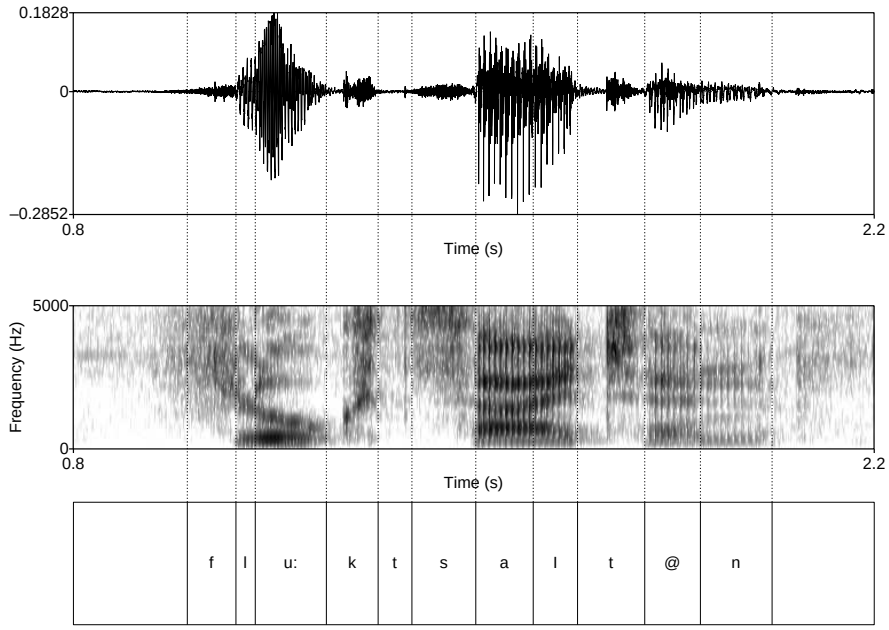

Figure 2.2.: Speech signal and spectrogram of the German word ”Flugzeiten”.

[flu:ktsaIt@n] (flight time).

As can be seen, there exist frequency ranges which exhibit a particularly high energy over the duration of certain segments. For example, four such high energy frequency bands can be observed in the segment [u:]. These characteristic energy peaks only occur in voiced phones and are especially pronounced for vowels. They are called *formants*. In order to differentiate between the different formants they are numbered according to their place of occurrence. Thus, the formant with the lowest frequency is called the first formant or F1, the formant with the next higher frequency is the second formant, F2, and so on. Generally, up to five formants can be observed in a voiced stretch of speech depending on the frequency range displayed. However, the most important formants that distinguish between different vowels are the first two formants F1 and F2. It has been observed that especially the first two formants can be related to certain articulatory movements.

For example, a low first formant frequency indicates a high tongue position as in [i:] and [u:]. A high F2 value indicates a fronted tongue position as in [i:]. The rounding of the lips can generally be observed by lower frequencies of F2 and F3. These effects are nicely shown in the vowel triangle of the German cardinal vowels [a:] [i:] [u:] (cf. Figure 2.3). While [i:] and [u:] share a low F1 which indicates a high tongue position the [a:] represents the vowel with the highest F1 value indicating a very low tongue position. Indeed,

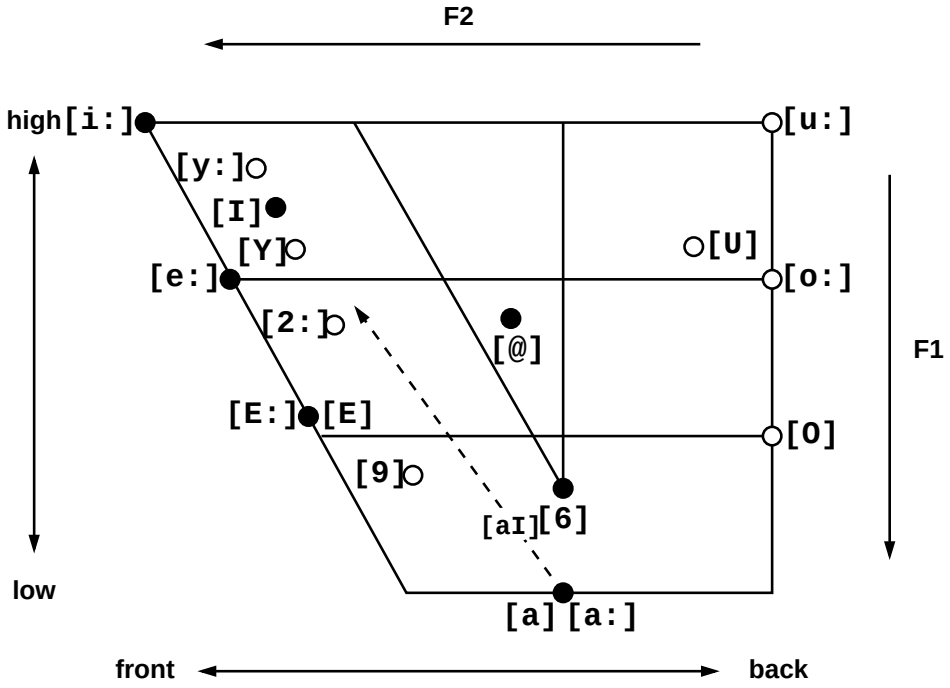

Figure 2.3.: Vowel space showing the German vowels. The vertical axis displays the articulatory high-low dimension as reflected in the F1 values. The horizontal axis represents the front-back dichotomy which is related to the F2 values. Rounded vowels are represented with an unfilled circle, unrounded vowels with a filled circle.

[a:] is the most open vowel where the tongue does not minimise the area for the airflow in the oral cavity The average F2 value for the German [a:] indicates that the tongue is relatively neutral in its position. In other languages there exist more phonemes which are distinguished by a more fronted or backwards tongue position. In these languages the vowel space is a quadrangle rather than a triangle[2].

Since the configurations of the vocal tract can be related to static acoustic characteristics, the movements of the articulators can be related to movements of the formant frequencies [Ste99, pp. 472-477]. For example, the diphthong [aI] is represented in the vowel space as a path starting at [a] and moving towards [I][3] . This corresponds to the movement of the tongue from a low to a high position as represented by the lowering of the F1 frequency while the F2 change indicates that the tongue is moved towards the front of the oral cavity.

[2] For a more detailed overview of vowel systems see [CY90, pp. 62-72].

[3] Although diphthongs are generally interpreted as one phoneme the SAMPA subset used for speech recognition experiments does not contain diphthongs. In those experiments diphthongs are modelled as two vowels in order to avoid sparse data problems which can occur when additional phonemes are defined. However, phonetically diphthongs are treated as one phoneme.

This change can be observed in the spectrogram in Figure 2.2. While the first two formant frequencies for the [a] lie relatively close together they shift apart for the [I].

For consonants such an articulatory-acoustic relationship can also be established. However, since consonants generally do not exhibit a formant structure the acoustic cues for different manners of articulation are related to the overall spectral envelope.

Consonants can be divided into sonorants and obstruents. For sonorants the degree of constriction on the airflow is higher as compared to vowels but lower as compared to obstruents where a complete occlusion takes place. Sonorants are characterised by a continuous phase as can be seen in vowels. For example, the part of the spectrogram which corresponds to the [s] shows an almost constant energy distribution in the higher frequency ranges while there is almost no energy in the lower frequencies. Since the [s] is a voiceless sound no formants can be observed. In contrast, for the obstruent [t] there is a phase of closure where almost no energy at all can be observed. The release of the closure can be deduced from the a short phase of turbulence characterising the burst which is sometimes followed aspiration. Only then an increase in the energy level of the formant frequencies of the following vowel occurs. This is a very characteristic structure for plosives. There exist many detailed analyses on the acoustic characteristics of certain articulatory features. For a more detailed survey of articulatory-acoustic relationships see [Ste99].

2.1.4. Sources of Variation

These systematic relationships between articulation and acoustics suggest that the acoustic signal of a word or a phone ought not change significantly when produced by different speakers in different situations. However, there are many sources of variation not only between different speakers but also within the articulations of one speaker which introduce considerable noise in the realisations of phones and words.

One of the most severe causes of variation is coarticulation. Although it is generally assumed that phones exist as independent entities in the articulation plan, articulation is a continuous process which involves the movements of more or less slow articulators such as the tongue body, the tip of the tongue or the jaw. It is, therefore, impossible to produce a discrete series of phones independently of each other. It is rather the case that consecutive phones are linked to each other by transition phases where the articulators move from one position to another. This leads to what is sometimes called ”sloppy articulation”: the articulators already begin to move towards the following phone while the current one is not yet finished.

Coarticulation can occur in different situations. For example, an articulator which is not necessary for the production of the current phone can begin with the movements for

the following phone. In the production of the plosive [t] the lips are not involved in the articulation. In the sequence [to] the lips can therefore start with the rounding while [t] is performed. The result is a rounded [t]. In contrast, in the case where an articulator is necessary for both phones a stronger influence of the articulation of both phones is exerted. For example, in the sequence [tj] the tongue tip is retracted for the [t] in anticipation of the following approximant [j] where the tongue body has to reach the palatum. Similarly, the place of articulation of the [j] is shifted towards the front. Thus, the place of articulation of both phones is altered due to coarticulation.

Apart from the place of articulation other features such as the type of articulation or voicing can also be altered. In the sequence [tj] both phones can be affected even more because the voiceless [t] is likely to have a devoicing effect on the [j] which would result in [tC]. However, it might also be the case that the [t] becomes voiced due to the influence of the voiced [j]. In this case the sequence would more resemble [dj]. It is difficult to predict which kind of coarticulation will take place. However, there is a common agreement that overall anticipatory coarticulation is predominant where the actual phone is affected by a phone still to come [CY90, p. 123].

Sometimes coarticulation is used synonymously with *assimilation*. However, some authors distinguish coarticulation from assimilation (e.g. [Woo96]). According to this point of view assimilatory processes are assumed to take place at a higher processing level in the speech production. They are characterised by language specific constraints. Assimilations generally cause allophonic phoneme variations which can be captured within the basic phonetic transcription inventory. For example, the English prefix *in* /In/ becomes /Im/ before bilabial consonants as in *impossible*. The /n/ remains unaltered before non-bilabial consonants as in *intolerable* or before vowels as in *inactive*. Thus, it is assumed that in the process of speech production it is already planned to articulate an /m/ instead of an /n/ in the word *impossible* which can even be reflected in the orthographic representation of the word.

In contrast coarticulation is assumed to be caused by the physical restrictions of the articulatory system alone and is not planned. In an attempt to distinguish both notions Wood [Woo96, p. 139] defines assimilation as a ”contextually determined and language-specific allophonic variation of a subset of phonemes” while coarticulation is rather a ”local articulatory adjustment of all phoneme instantiations to their current neighbours”.

This implies that coarticulatory processes are assumed to be universal and to occur similarly in every language. However, this assumption is not unanimously shared in the literature. For example, Kohler states that coarticulation does not have to occur in one physiologically predetermined way but can follow different strategies in different languages [Koh77, p. 89]. Some authors therefore conclude that there is no difference between assimilation and coarticulation [PW83] by arguing that the so called physiologically caused coarticulation is in fact a planned process.

In this work no attempt will be made to distinguish between assimilation and coarticulation. Since the distinction is based on the notion of a ”planned process” during speech production it would be difficult to follow this discrimination in an investigation of a large corpus of spontaneous speech. Therefore, in the further process of this work only the notion of *coarticulation* will be used which comprises both, phenomena which are sometimes classified as coarticulation and those classified as assimilation.

Apart from coarticulation further variation of the speech signal is introduced by individual physiologies of speakers. On the one hand, the fundamental frequency can vary considerably from speaker to speaker because of the different sizes of the vocal folds. Since this only affects the source but not the filter of the articulation process one might argue that this does not affect the specific spectral characteristics of the different phones. However, the sizes of the vocal tracts of different speakers are also highly variable. This effect is especially strong between the vocal tracts of adult female and male speakers. Even more, the typical adult male vocal tract is not simply a scaled-up version of a typical female one, but it is disproportionately longer in the pharynx. This means that the relationship between the formant frequencies of female and male speakers is non-linear. In fact, the formant configurations are determined by a particular size relationship between different parts of the vocal tract. Since these size relationships vary among different speakers the formant frequencies of these speakers are bound to exhibit slightly different configurations while the overall vowel chart remains similar [Nol99, p. 750].

Speakers can also differ in their individual articulation strategies. For example, it can be observed that speakers use different strategies when speaking faster. While some speakers tend to move their articulators faster other speakers simply reduce the movements and produce a more ”sloppy” articulation. It could be shown that there exists a relationship between the kind of strategy that is applied and some physiological features such as the tongue or jaw size [KM76]. However, the physiology of a speaker can only explain some of the variations found in the speech signal. There are many other factors which cause significant differences in articulation and the acoustic characteristics of the same phone or word.

To sum up, there exist systematic relationships between articulation and acoustics. However, variations of phonetic segments occur to a large degree in the speech of one speaker due to coarticulatory processes. Inter speaker differences occur to a large extent due to physiological differences of the articulatory system but also due to individual articulation strategies. Thus, despite the systematic relationships between articulation and acoustics considerable noise is introduced into the signal in the process of articulation.

2.2. Automatic Speech Recognition

While in acoustic phonetics articulation is directly related to specific acoustic characteristics, in standard approaches to automatic speech recognition the speech production and articulation processes are interpreted as variables of a stochastic process. The goal of these types of speech recognition is to deduce from the acoustic signal the words that have been spoken given knowledge of the statistical relationship between the signal and the spoken words.

This statistical relationship is described in the *channel model* (cf. Figure 2.4) where the different speech production modules are described in terms of probabilities. In this model the complex process of formulating a linguistic message consisting of a sequence of words is represented by its observation probability $P(w)$.

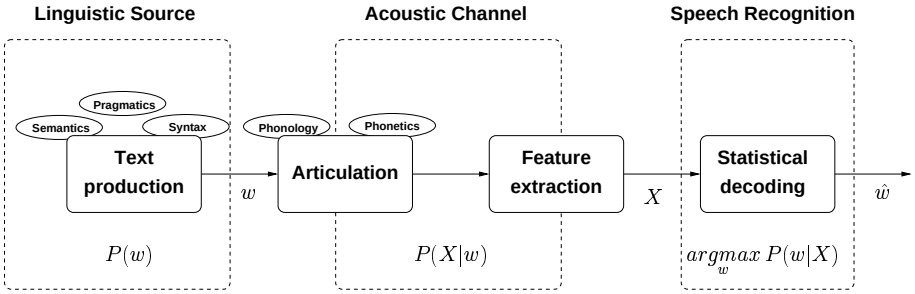

Figure 2.4.: Probabilistic model of speech production and speech recognition.

The acoustic characteristics of the speech signal which are influenced by variations that occur in the process of articulation are part of the next variable describing the acoustic channel. As has been shown, during articulation variations can be caused for example by different speaker specific physiologies, individual pronunciations, dialectic variants etc. Further sources of noise which affect the acoustic speech signal are introduced in the transmission of the signal from the articulators to the final representation in terms of feature vectors. This means that the environment of the recordings plays an important role. For example, recordings in a relatively quiet office environment have a different effect on the speech signal than a noisy environment. Also, different microphones introduce different effects. Finally, features are extracted from the received speech signal. This is an important step in the signal processing procedure because it decides which characteristics will be represented and which will be ignored.

In the channel model this process from the intended word sequence to the observed feature vectors is represented by a variable which denotes the probability $P(X|w)$ of observing the feature vectors X given the word sequence w.

The goal of speech recognition is now to find the ideal word sequence w^* that has pro-

duced the observed sequence of feature vectors. Since in the production process different sources of noise are introduced[4] there exists no unique solution to this problem. In the probabilistic approaches to speech recognition this problem is therefore rephrased to the search of a word sequence $\hat{w}$ which maximises the probability $P(X|w)$. In other words, this word sequence $\hat{w}$ has to meet the following condition:

$$\hat{w} = \underset{w}{argmax}\, P(w|X) \tag{2.1}$$

In order to determine $\hat{w}$ it is therefore necessary to compute $P(w|X)$. This can be done by decomposing the probability into its components according to Bayes' rule:

$$\underset{w}{argmax}\, P(w|X) = \underset{w}{argmax}\, \frac{P(w)P(X|w)}{P(X)} \tag{2.2}$$

Since $P(X)$ represents a constant value with respect to the maximisation it can be neglected in the search for the word $\hat{w}$ with the highest value for $P(w|X)$. Thus, it is sufficient to find the word sequence $\hat{w}$ which complies with the following requirement:

$$\hat{w} = \underset{w}{argmax}\, P(w)P(X|w) \tag{2.3}$$

This means that the optimal word sequence $\hat{w}$ can be determined only by considering the production variables of the channel model which consists of the linguistic source, described by $P(w)$, and the acoustic channel as represented by $P(X|w)$.

There exist several approaches to modelling the probability of a certain word sequence $P(w)$. In speech recognition, probabilistic *language models* have become a standard method for the estimation of this probability. They estimate the probability of one word given one or more preceding words. The more words that are taken into consideration for the estimation of the probability of the next word, the more detailed the language model is, but the more training data is needed in order to reliably estimate the probability of each possible word chain. Therefore, generally, so called *bigrams* are used, where the length of the word chains upon which probabilities are estimated is restricted to two. However, since the current work focuses on the acoustic aspects of speech rate variation the language model will be neglected

[4] It should be noted that there are many systematic aspects of production causing characteristical acoustic effects. However, in the statistical approaches of automatic speech recognition they generally appear as noise.

in the further course of this work. But it should be noted that the results of a speech recognition system are always influenced by both the language model and the acoustic model.

In order to model the acoustic channel which is described by the probability $P(X|w)$ it is necessary to provide an acoustic model for all words w. However, before such a model can be established it is necessary to define features with which the relevant characteristics of the speech signal can be represented.

2.2.1. Feature Extraction

From the preceding section one might expect that formants are good features for speech recognition. However, formants exhibit certain characteristics which make them less suitable for this task. The most obvious argument against formants is the fact that they can only be computed on voiced speech. This means that for voiceless consonants no formant information can be derived from the signal, which cannot be tolerated for the task of speech recognition. Another drawback is that the absolute values of formants are highly speaker specific [LB52]. Not only do the fundamental frequencies of female and male speakers differ significantly but it is also the case that the specific vocal tracts exhibit different characteristics which are reflected in a non-linear transformation of the formant frequencies of different speakers [Nol99, p. 750]. Finally, automatic formant tracking is still a difficult task that is highly error prone especially for recordings made in noisy surroundings. For speech recognition it is desirable to have features that are easy to compute and less dependent on speaker specific vocal tracts.

For these reasons the so called *Mel-Frequency Cepstral-Coefficients (MFCC)* are used as the standard features for the representation of the speech signal in speech recognition. They are computed over a certain stretch of speech in order to capture frequencies from around 60 Hz to over 8,000 Hz. Generally, around every 10 ms a new feature vector is computed. This period is often referred to as the *frame rate*. It represents the scope of the speech signal to which the classification of the feature vector is aligned.

The MFCCs are based on the so called *mel-spectrum* which is derived from the spectrum by summing the energy of higher frequencies while retaining the resolution of the lower frequencies by the application of a mel-filter bank. The filter bank consists of overlapping filters which sums the energy of certain frequency bands (cf. Figure 2.5). With this procedure the higher sensitivity of the human ear to lower frequencies is modelled. The effect is a non-linear scaling of the frequency axis which is similar to the mel scale. In general, the mel-spectrum shows a high resolution in the frequency range up to the second and third formants.

From this mel-spectrum the cepstral coefficients are derived by a transformation which

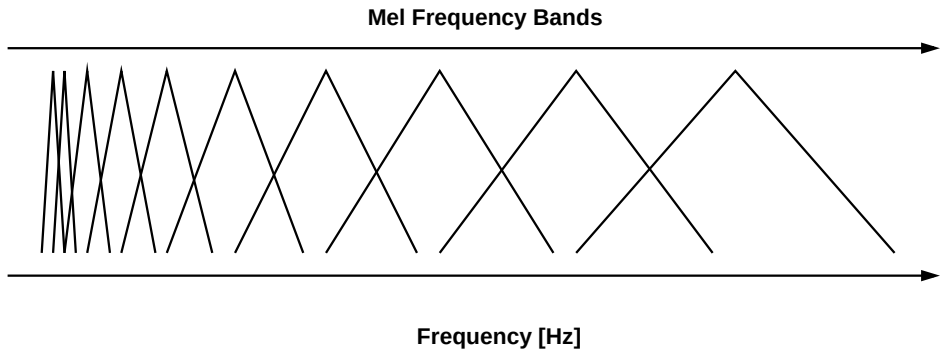

Figure 2.5.: Mel-filter bank consisting of overlapping filters which pools the energy of certain frequency bands.

analyses the broad and fine structure of the mel-spectrum. Since this transformation can be interpreted as performing a spectral analysis of the spectrum the result is referred to as *cepstrum* which represents an anagram of the term ”spectrum”. The lower coefficients of the cepstrum give a representation of the broad structure of the whole spectrum while the higher coefficients capture effects of the fine structure which consists mainly of the harmonics. Since the harmonics do not convey information about the identity of the phone the higher coefficients are discarded in the final feature vector. Thus, only the lower cepstral coefficients are kept for the further processing. Since it can be shown that the cepstral coefficients of a spectrum are de-correlated [ML93] it is assumed that the MFCCs are also roughly de-correlated and constitute a suitable base for the feature set in a classification task [HAH01, p. 64].

In order to integrate information about the dynamic characteristics of the speech signal over time, the first and second order derivatives of the MFCCs are incorporated into the feature vector. They are generally computed over several consecutive frames which capture the dynamics over about 50 ms of the surrounding signal.

In addition to the MFCCs and their derivatives, the overall energy of a frame of speech is computed and added to the feature vector. Since the overall energy provides information about voicing this is a further valuable feature.

2.2.2. Acoustic Modelling

In order to set up acoustic models of words that are based on these features it is necessary to find a suitable framework which is able to represent not only the spectral characteristics as provided by the feature vectors but also their changes over time. Also, more suitable entities than words have to be derived because models of whole words require many train-

ing instances in order to estimate a reliable set of parameters. This is especially important for a large vocabulary. Therefore, smaller units which occur more often have to be derived. Such appropriate units are, for example, phones. As has been shown in the previous section words can be represented by a phonetic transcription which has a close relationship to the acoustic and articulatory characteristics of the spoken word. However, due to coarticulatory influences the realisations of a phone are strongly dependent on its context. In order to capture such coarticulatory variations phones are therefore defined in relation to their context which is generally restricted to one phone on each side. Such context dependent phones are called *triphones*. Thus, for one phoneme there exist many different triphones which represent a phone with different coarticulatory influences. For example, the German word "nein" `[naIn]` would be represented by the following sequence of triphones:

nein := `#/n/a n/a/I a/I/n I/n/#`

In this example `n/a/I` denotes the phone `[a]` but in order to distinguish it from realisations of `[a]` with a different coarticulatory influence the context is added to the notation. Thus, `n/a/I` represents an `[a]` with the left context being an `[n]` and the following phone being `[I]`[5]. After finding a suitable unit for the representation of words the question arises what kind of framework is able to model the acoustic characteristics and to capture the changes over time.

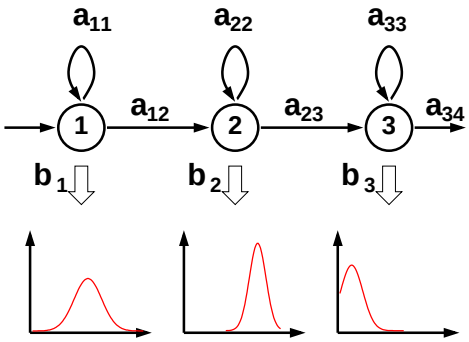

Figure 2.6.: Hidden-Markov model consisting of three states which represent the triphone `n/a/I`. The figure illustrates one dimensional feature vectors representing the first formant modelled by a single Gaussian distribution.

In automatic speech recognition so called *Hidden Markov Models (HMMs)* prevail, a probabilistic modelling strategy that is able to meet these requirements. Formally, an HMM is a finite automaton which models a two-stage random process. The first random process

[5] If the phoneme `[aI]` was treated as one segment there would be only one vocalic triphone `n/aI/n` in the triphon representation of the word. However, as has already been pointed out earlier, in many automatic speech recognition systems diphthongs are treated as two distinct phonemes.

describes the probability of being in a certain state of the model. The second process represents the probability of observing a certain feature vector while being in this state. This means that an HMM consists of states which are connected by transitions that model the durational characteristics of a sub word unit. In order to capture the spectral characteristics they are also provided with the ability to emit feature vectors with a certain probability. The assumption behind this model is that the production of a sequence of feature vectors can be modelled by a Markov model. Since only the feature vectors can be observed the ”real” state sequence which produced the observed feature vector sequence remains hidden. Hence the name ”Hidden” Markov Model. The problem of recognition is thereby reformulated into finding the corresponding state sequence given a sequence of observed feature vectors. If the state sequence that produced the signal is known the identity of the underlying word is also known because usually every state represents a triphone.

The acoustic properties that belong to a certain state are modelled by probabilistic parameters by means of Gaussian distributions. Since this representation provides a continuous output distribution such models are called *continuous* HMMs. The distributions describe the probability density of a certain feature vector being emitted in this state. For example, Figure 2.6 shows schematically the emissions of the states that model the triphone `n/a/I`. For the sake of illustration the feature space in this example consists only of one dimension representing the first formant. The example demonstrates that the outer states of a triphone are supposed to model coarticulatory influences of the neighbouring phones. In this case the first formant’s frequencies modelled by the first state show a flat distribution in a rather neutral place. This indicates the influence of the neutral context `[n]`. The middle state is supposed to model an ideal un-coarticulated part of the `[a]` with a high first formant. The last state shows finally the influence of the context vowel `[I]` with a low value for F1.

However, this is a very simplified example. Normally, the output probabilities are modelled by more than one Gaussian distribution, where every distribution is weighted by a parameter. In order to reduce the number of parameters it is possible to share the Gaussian distribution between all states. In this case a codebook contains all distributions which are valid for all states. The states only retain the weights for each distribution. Such models are called *semi-continuous* HMMs.

HMMs also provide the possibility for a more flexible duration modelling by permitting different kinds of topologies. However, in general two topologies prevail. The example in Figure 2.6 shows a linear topology where from every state either a loop can be performed or a transition can be made into the next state. In order to provide more flexibility in the modelling of duration so called *Bakis models* can be applied (cf. Figure 2.7). In addition to the linear structure they permit the following state to be skipped and the next-but-one state to be entered directly.

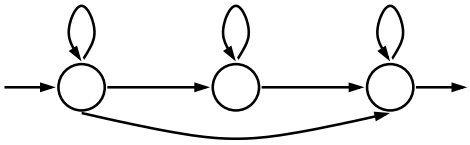

Figure 2.7.: Bakis model as an alternative HMM topology for a more flexible duration modelling.

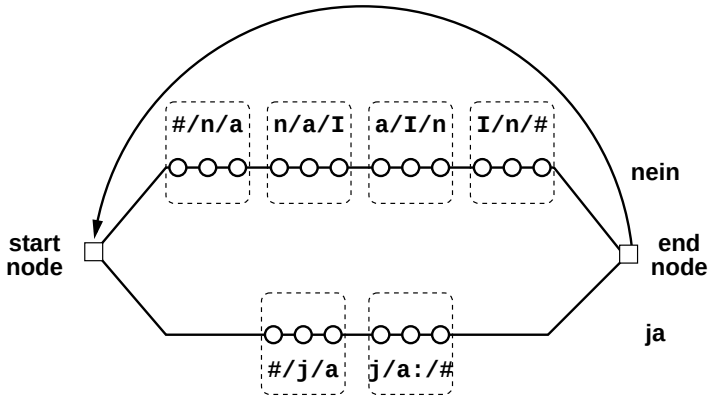

Figure 2.8.: Linear structure of the lexicon containing words composed of triphones HMMs.

To sum up, an HMM is a finite automaton ϕ which is defined by a triple of parameters:

$$\phi = (A, B, \pi)$$

This means that an HMM ϕ is defined by a matrix A containing the transition probabilities a_{ij} for going from state i to state j and a set B of parameters containing the output probabilities $b_j(x)$ of observing the feature vector x while being in state j. This probability is generally represented by several Gaussian distributions in order to model multi-modal distributions. Therefore, each state j contains a set of Gaussian distributions which are called *mixture densities*. Each distribution g_{jk} is defined by a mean vector μ_{jk} and a covariance matrix K_{jk} and has a corresponding weight c_{jk}. For semi-continuous HMMs there exists only one set of Gaussian distributions g_k which is valid for all states. In this case the output distributions of each state j consist of the weights c_{jk} for the corresponding state independent Gaussian distribution g_k. In order to model the initial transitions into the first state of a model a vector π is defined which contains the probabilities of starting with state i.

By concatenating such basic triphone HMMs whole word models can be created which represent the lexicon by a simple parallel architecture as shown in Figure 2.8. In this example the lexicon consists of the two words ”ja” (yes) and ”nein” (no) which are modelled by the corresponding triphones. Thus, each word is represented by a sequence of states. As can be

seen each word starts from the same point and ends in the same ending point. These start and end points are pseudo states with no emission probabilities and only serve as a means to bundle the paths after a word end. Generally, this representation of the lexicon is transformed into a more efficient tree organisation where identical pre-fixes are shared by different words. At the end of one word the starting point can be reached via a transition from where the next word can be accessed.

2.2.3. Decoding

Given the acoustic model of the words the goal of decoding is now to determine the state sequence s^* with the highest probability of producing the observed vector sequence X. Once the state sequence is known the identity of the word is revealed, thus the word chain "recognised".

In order to determine the probability of a path, the state space is expanded by aligning the observed feature vectors with the state sequence. The path probability is obtained by computing the transition and output probabilities given the aligned feature vectors. The most intuitive approach to determine the most probable state sequence would be to expand the whole search space by computing the probabilities of all state sequences that can possibly have generated the observed sequence of the length T. Since the model consists of a probabilistic approach, any state sequence of the length T can generate the observation sequence, however low the probability may be. Depending on the size of the lexicon and the length of the observation sequence this means an intolerably high computational load.

A more efficient procedure is provided by the *Viterbi algorithm*. It reduces the computational load by discarding paths that lead into a state s_t which is also reached by a different path with a higher probability. This is achieved by finding partially optimal paths. Thus, at the last time step T the locally optimal path represents the global best path and determines the recognised word hypothesis. Generally, this process is further optimised by a beam search strategy where paths with a low probability are excluded from further processing.

While the Viterbi algorithm is a standard technique in automatic speech recognition other efficient algorithms for different recognition tasks have been developed. However, in the current investigations only the Viterbi algorithm is used for decoding.

2.2.4. Training

The decoding process requires a model with parameters that optimally fit the observed data. This means that the model should produce a maximal probability $P(X|w)$ for an observation sequence X that belongs to the word w. No method is known that finds the optimal

parameters for a given training sample. However, this modelling approach allows an iterative optimisation of the model parameters where the probability of the model to produce the observation sequence X is higher for the optimised model $\hat{\phi}$ than for the old model. The training implies that an initial model does exist. The problem of initialisation is generally solved by another probabilistic approach that estimates the initial Gaussian distributions from the training data.

A standard technique for the re-estimation of the model parameters is the *Baum-Welch algorithm*. It estimates the new parameters for a state from the observed feature vectors by weighting the observed values with the probabilities of being produced by that state of the model ϕ.

For example, the probability $\hat{a}_{ij}$ of making a transition from state i into state j is determined by the expected number of transitions from i to j divided by all expected transitions from i into any other state given the observed sequence X of feature vectors:

$$\hat{a}_{ij} = \frac{\text{expected number of transitions from } i \text{ to } j}{\text{expected number of transitions from } i \text{ into any state}}$$

For this computation it is necessary to know how probable it is to be in state i at any given time step t and making a transition to state j given the observed vector sequence and the model ϕ. Let this probability be denoted by the variable $\gamma_t(i,j)$ and the probability of being in state i at all be denoted by $\gamma_t(i)$. Then the new transition probability $\hat{a}_{ij}$ can be derived according to the following formula:

$$\hat{a}_{ij} = \frac{\sum_{t=1}^{T-1} \gamma_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \tag{2.4}$$

Thus, the expected number of transitions from state i to state j is determined by the observed number of transitions from state i to state j weighted by the probability of making that transition given the observed vector sequence X and the old model ϕ.

The output probabilities are re-estimated according to the same principle. In order to determine the new mean of a Gaussian distribution k by all feature vectors x_t the probability $\zeta_t(j,k)$ of being in state j and selecting the k-th distribution given the observation X and the model ϕ has to be computed first. The new mean $\hat{\mu}_k$ is then determined as the sum over all vectors x_t multiplied by the probability $\zeta_t(j,k)$ divided by the sum over all time steps t of $\zeta_t(j,k)$:

$$\hat{\mu}_{jk} = \frac{\sum_{t=1}^{T} \zeta_t(j,k) x_t}{\sum_{t=1}^{T} \zeta_t(j,k)} \tag{2.5}$$

The new values of the covariance matrices $\hat{K}_{jk}$ and the weights $\hat{c}_{jk}$ of the Gaussian distributions are computed accordingly by making use of $\zeta_t(j,k)$.

To sum up, in order to derive new model parameters that optimise the probability of observing X given the model ϕ it is necessary to provide an initial model ϕ and a training sample containing the training vector sequence X. The new parameters are estimated by weighting the observed data with the probability of their occurrence according to the old model.

2.2.5. Evaluation of the System

In order to evaluate the performance of a speech recognition system it is important to provide a meaningful measure. The most common measure is the *word error rate (WER)*. As the name already indicates it is based on the number of errors in the recognition result at the word level. For this measurement the word hypotheses produced by the speech recognition system are compared to the correct word sequence as provided by a correct orthographic transcription of the test utterance. Since the recogniser does not necessarily produce the same number of words as in the correct word sequence an alignment is applied where the correct word sequence is matched with the word hypotheses. This alignment is performed in such a way that the words which are correctly recognised match the corresponding words in the reference transcription. From this alignment the number D of deletions and the number I of insertions can be determined. When a wrong word has been recognised this is counted as a substitution S. From these counts the WER is determined according to the following formula:

$$\text{WER} = 100 \cdot \frac{S + D + I}{W}$$

where W denotes the number of words as provided by the reference transcription of the utterance. Thus, the word error rate denotes the percentage of incorrectly recognised words.

Generally, improvements of the performance of a system are measured in terms of the relative reduction of the WER. This allows changes over tasks with different levels of complexity to be compared, since absolute improvements on an already well performing system or corpus are generally smaller than on systems with a lower performance.

2.3. Comparison of the Acoustic Features

This brief introduction into the basic methods of phonetic investigations and automatic speech recognition shows that there are discrepancies in the representation of the acoustic

characteristics of speech between both disciplines. As has become obvious, the most important paradigm for the description of the acoustic characteristics of vowels in phonetics is the use of formants. Formants seem to be an appropriate means for describing acoustic events because they relate so well to articulation. On the other hand, in automatic speech recognition formants are not used, because they lack certain characteristics demanded by classification approaches. Apart from carrying speaker specific information which has to be normalised, formants can only be computed on voiced speech and are difficult to track automatically. In contrast, MFCCs can easily be computed automatically on any part of speech, even on silences or non-speech events, and capture characteristics of the whole spectrum. Furthermore, the MFCCs are able to separate information of the fundamental frequency, as reflected in the harmonic structure of the spectrum, from the information of the vocal tract configuration which is important for the identification of a phoneme. However, they are far more difficult to interpret in their relation to articulation than formants. In order to project results of phonetic analyses based on formant frequencies onto automatic speech recognition, one has to address the question of the relationship between the two sets of features first.

In the literature many investigations can be found on how to capture characteristic information of vowels from the spectra. A very interesting series of experiments is reported in [PPvdG67] [PvdKP69] [KPP70] [PTP73]. These experiments are based on a principal component analysis of the spectrum which is compared to results obtained by formants. For the experiments, the static part of vowels taken from one-syllabic words uttered in isolation was analysed.

A representation of the vowels by the first two principal components of the logarithmically spaced spectrum was computed which established a new topology of the vowel space. A comparison of this representation with the vowel space given by the first two formants reveals that both representations are in fact congruent. This means, that the first two formants of the steady state part of a vowel can be predicted by a principal component analysis of the spectrum.

For the comparison of formants and MFCCs this indicates that MFCCs and formants are also very similar. In the literature it is generally agreed upon that the MFCCs are in fact decorrelated [HAJ90, p. 64]. This indicates, that the inverse cosine-transform, which is applied to the mel-filtered spectrum, performs some kind of principal component analysis. Thus, the principal components of the spectrum as reported in [PPvdG67] can be assumed to be quite similar to the MFCCs. This would mean that the first two formants and the first two MFCCs are closely correlated, at least on static vowel segments.

However, differences between formants and MFCCs were found in [ZJ93]. In a comparison of formants and MFCCs it was shown, that MFCCs are more appropriate for the classification of vowels than formants because they provide a more complete representation of the vowel spectrum. Again, experiments were carried out on vowels taken from CVC

syllables spoken in isolation. In this investigations the first MFCC is interpreted to capture the overall spectral tilt while the second coefficient is a measure of the spectral compactness. The higher coefficients are then representations of more and more finer details of the spectrum. In a classification task it was shown, that using the first three formants as a representation of the vowels yielded slightly worse recognition rates than when using the first 10 MFCCs. However, when comparing the three formants with the first three MFCCs the formants significantly outperformed the MFCCs. This indicates, that the few formants convey more information than the corresponding MFCCs. Thus, there is no one-to-one relationship between formants and MFCCs. When adding the fundamental frequency as additional source of information to the three formants and the 10 MFCCs, the performance of the formants increases significantly while the results of the MFCC features do not change. However, the formants never reached the performance of all 10 MFCCs. In summary, this investigations showed, that the classification performance of formants and MFCCs differs significantly in favour of the MFCCs. A direct comparison of the first three coefficients of each feature set also shows, that MFCCs and formants can not be directly compared.

To sum up, formants and MFCCs show a similar behaviour in a classification task. Due to the higher number of coefficients MFCCs seem to convey more information that is necessary for a classification. Thus, in a classification task MFCCs and formants behave comparably but not exactly similar. There are also indications that MFCCs reflect the vowel topology as observed in the representation of the first two formants. Thus, although formants and MFCCs are clearly distinct features there are strong indications that they reflect similar characteristics of the speech signal.

3. Influence of Speech Rate on Acoustic-Phonetic Properties of Speech

The variation of speech rate has been considered in a variety of acoustic-phonetic experiments. The effects found in these experiments are manifold and do not only concern durational and spectral parameters but also affect the human perception of speech. This chapter reviews the influence of speech rate as one of many factors on acoustical and articulatory parameters as well as on perception.

3.1. Durational Effects

Speech rate is a complex variable which is composed both of the rate at which the speech itself is produced — the articulation rate — and the number and duration of pauses in the utterance — the pause rate [MGL84]. *Articulation rate* measures the number of units per time that are produced by the articulators. It is strongly constrained by the physics of the articulatory system and cannot exceed a certain rate. Generally it is measured in terms of syllables per second on stretches of speech where pauses are discarded. Closely related but not to be confused with is the *articulator speed* which gives the speed, with which the articulators move during speech production. Although one might assume that a fast articulation rate always causes fast movements of the articulators, this is not necessarily the case. As will be shown in the following section on spectral effects of rate variation, it is possible to preserve a slow articulation speed in speech with a high articulation rate by a more sloppy execution of the required movements.

In the following section a distinction will be made between *articulation rate*, which is measured only on stretches of speech where silences are discarded, and *speaking rate*, which means the overall rate where silences are included. The term *articulator speed* refers to the velocity of the movements of the articulators and is different from the articulation rate. If no distinction is intended the general term *speech rate* will be used.

In early investigations the change of the overall speech rate was attributed to changes in the pause rate while the articulation rate was considered to be a speaker specific constant. This would mean that a slow speech rate is characterised by more and longer silences while

a fast rate is achieved by simply shortening and omitting pauses. The actual time without pauses during which speech is produced would remain the same. In [GD75] conversational speech from radio interviews was analysed by computing the number of syllables per second over a certain stretch of speech. In order to find appropriate stretches of speech, the interviews were divided into so-called *runs* which occur between two pauses. Several consecutive runs were grouped together to form a stretch of speech containing around 30 syllables. Over these groups of 30 syllables the articulation rate was measured by computing the mean number of syllables per second. However, the duration of the pauses was not taken into account. Over all interviews the pause rate varied by about 27% while the articulation rate only varied by about 10% which was considered a marginal amount. The authors concluded that most of the speech rate variation is caused by changes in the pause rate while the articulation rate remains relatively stable for each speaker and constitutes a speaker specific constant.

This argumentation was not followed by the authors of a reanalysis of this data [MGL84]. They argued that the stretch of speech over which the articulation rate was measured was too long. Instead of a group of 30 syllables they computed the articulation rate over shorter runs, which occur between two pauses. The mean number of syllables of these runs was around 11. With this new base of the measurements the articulation rate varied from 2.2 to over 7.9 syllables per second over all speakers with mean syllable durations between 126 and 450 ms. The authors concluded that the articulation rate does indeed vary over a wide range even for one speaker and that the variation of the overall speaking rate consists of both changes in articulation rate and in pause rate.

From these experiments the conclusion can be drawn, that the articulation rate is indeed severely affected by overall speech rate variations. For a review of the acoustic implications it is therefore more interesting to focus on variations of the articulation rate, because the effects of rate variation on the articulation will be reflected in the acoustic characteristics of the speech signal.

While the above mentioned results show that too long a stretch of speech for the measurement of the speaking rate is not appropriate for measuring local rate variations, it should be mentioned that this stretch of speech must not be too short either. The shortest entity for measuring the syllable rate would be the syllable itself. However, since the syllables of a language can have different structures this would not be a very reliable measure. For example, a syllable can consist of a single vowel as the `[a]` in the German verb *reagieren* `[re|a|gi:|r@n]` (to react). It is obvious that the duration of this syllable will be significantly shorter even in slow speech than the complex syllable `[hE6pst]` in German *Herbst* (autumn). In a language that allows complex consonant clusters such as German such differences are likely to appear.

One might argue that the phone rate would be a better measure since it represents a less heterogeneous structure. However, phones exhibit intrinsic durations. For example, the En-

glish vowels can roughly be divided into two categories, for which duration is one distinctive feature, namely the intrinsically short and long vowels. Since the duration of vowels is also affected by consonantal context and stress, high local variations of the phone rate can be expected due to these factors alone, rather than to articulation rate. If the phone rate is smoothed over a longer stretch of time these variations are expected to be neutralised, leaving a rough estimate of the articulation rate. However, articulation rate is generally measured in syllables per second since the syllable is assumed to be a more adequate articulatory entity.

In a study on the relationship between duration and articulation rate over different stretches of speech it has been confirmed that in spontaneous speech a measurement of the articulation rate on stretches of speech where pauses are discarded is more appropriate [TKEB01]. The highest correlations between the number of linguistic units and the articulation time was found to occur on so called inter-pause stretches which occur between two pauses. In contrast, the correlation tended to be lower in intonation phrases, which could include pauses. Since intonation phrases can be longer or shorter than inter-pause stretches, so this result cannot be caused by a longer duration of the intonation phrases. In the same study the correlations of duration and segment rate computed on different speech units were investigated. As was to be expected, it was shown that the best correlation between the number of segments and duration was achieved for the shortest segments, the realised phones, while the longest unit, the intended word, only leads to minor correlations because words exhibit a higher variance in duration than phones.

With regard to articulation rate it is worth noting that different kinds of phones are affected differently by speaking rate. For example, vowels are affected far more than consonants. In [MTÁL97] a corpus with three different rates of speech was analysed. The Spanish corpus TRESVEL consists of utterances of speakers who were asked to speak slowly, fast or at a normal rate of speech. While the duration of vowels was reduced by about 61% from slow to fast speech, the reduction of unvoiced plosives only amounted to 36% which represents a relative difference of almost 100%.

In a more detailed analysis of the shortening of English vowels it was revealed that intrinsically long vowels are more affected by speech rate than short ones [Mil81], which reduces the absolute difference in duration between long and short vowels in fast speech. Apart from these phone-specific effects of the articulation rate, it has been shown that greater entities than a phone, such as the syllable, also behave differently under variations of the articulation rate. For example, unstressed syllables show a stronger shortening in fast speech than stressed ones [PL60] which actually increases the difference in duration between stressed and unstressed syllables.

To sum up, overall speech rate variation is generally attributed to both variations in the articulation rate, which affects the mean segment duration, and pause rate. However, it has been shown, that a measure of the articulation rate is always influenced by other factors,

such as the phone identity, the stress pattern or the syllable structure. In order to compensate effects of intrinsic duration, articulation rate is generally measured in syllables per second. The stretch of speech over which the measurement is taken should therefore on the one hand be long enough to compensate for different syllable structures. On the other hand it should be short enough to reflect the potentially high local variations of the articulation rate. The stress pattern remains a variable that is not captured by this measure. For acoustical analyses the articulation rate is the most interesting factor, because it affects the duration of the phonemic segments. In the following section it will be shown how these rather severe durational changes of the phonemic segments in spontaneous speech affect the spectral characteristics of speech.

3.2. Reduction

Early hypotheses about the influence of rate variation suggested that changes in the articulation rate simply cause a horizontal compression of the spectrogram where each sound segment is compressed in its duration while the formant frequencies remain unaffected (e.g. [Joo48]).

However, experiments showed that this hypothesis could not be sustained. In an investigation of the influence of duration on the spectral characteristics of vowels, Lindblom found evidence for *target undershoot* by measuring the formant frequencies of eight Swedish vowels uttered by one speaker in different stress and articulation rate conditions within the three consonantal contexts [bVb], [dVd] and [gVg] [Lin63]. He observed that with decreasing duration the formant frequencies of the tokens of any vowel became more heterogeneous. These variances were not only strongly correlated with the duration, but also with the consonantal contexts of the vowels. Based on these results Lindblom defined a function that predicts the frequencies of the n-th formant of a vowel token at the maximum or minimum of the formant track of a vowel. With this function Lindblom was able to explain about 50% of the variance of the formant frequencies observed in his data. The function is based on the distance between the hypothetical formant frequency that is characteristic for the consonantal context, the so called *locus*, and the *target*, i.e. the expected ideal frequency of the vowel. Since the locus depends on the consonantal context of the vowel, this distance is a measure of the expected coarticulatory strength of the context. The greater the locus-target distance, the stronger the target undershoot.

Lindblom estimated the target frequencies for each vowel by finding the optimal correlation between the amount of target undershoot — as defined by the difference between the observed formant frequency and the target value — and the duration of each vowel token.

The graphs for the Swedish vowels [I], [U] and [a] in the contexts [b-b], [d-d]

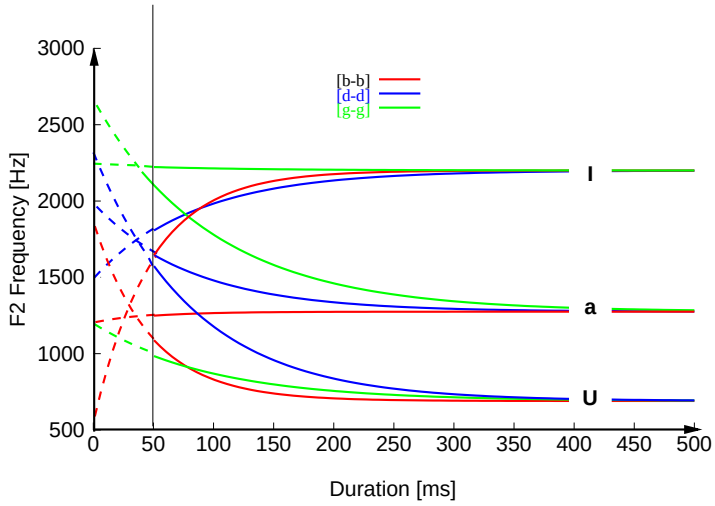

Figure 3.1.: Frequencies of the second formants of the Swedish vowels [I], [U] and [a] as a function of duration and consonant context as predicted by the target-undershoot model of Lindblom [Lin63].

and [g-g], as predicted by the target undershoot model, based on these target frequencies are shown in Fig. 3.1. As can be seen, the graphs for each vowel are approaching the same target frequency with increasing duration. However, as the dotted lines for durations of less than 50 ms reveal, this model is not realistic for very short durations. For example, the predicted formant frequency of an [U] in the [bVb] context for 10 ms is nearer to the target frequency of the [I] as the predicted value for [bIb], which in turn almost reaches the target frequency of the [U]. One would rather expect that the graphs for the same contexts intersect in one point at the hypothetical duration of 0 ms or less. These unrealistic predictions for very short vowels may be due to the fact that the recorded vowels had durations of at least 80 ms.

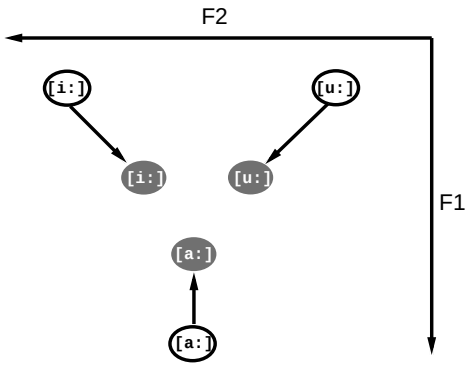

Figure 3.2.: Acoustically reduced vowel space as a consequence of target undershoot.

If the values of the predicted first and second formant frequencies of short vowels are plotted against each other a reduced vowel triangle as pictured schematically in Fig. 3.2 is obtained. While some authors state that this shows a tendency of short vowels to degenerate into the neutral vowel [@], Lindblom argues that this centralisation effect is simply the result of increased contextual assimilation and shows nothing about the dynamics of vowel articulation.

While the results of Lindblom's investigation show a strong relationship between vowel duration and a degeneration of the vowel formant frequencies, there exists no unanimous opinion in the literature as to whether duration is the only cause of this reductional effect. In the following section several other factors will briefly be described.

3.2.1. Causes of Reduction

Lindblom's observation of the target-undershoot phenomenon indicates a strong relationship between duration and the acoustic quality of vowels. However, one might also expect other factors to cause such a degeneration of the formant frequencies. As has been shown in the previous chapter, the formant frequencies are strongly correlated with the movements of the articulators. Therefore, target-undershoot can also be expected to occur in situations where less speech effort is applied during articulation. This can be the case, for example, in unstressed syllables. Thus, one might expect reduction to occur in unstressed syllables rather than in stressed ones. Furthermore, Lindblom's study shows that vowel reduction depends strongly on both vowel duration and the phonetic context of the vowel. This means that while the duration determines the extent of the reduction, the context determines the direction of the formant shift. This would explain Lindblom's observation of an increased variance for short vowels. However, if only the mean values of the tokens of a vowel were computed, no effect could be observed. In the following some findings about different causes of reduction will be briefly summarised.

Since duration and stress are strongly correlated, it can be argued that it is rather stress than duration that causes vowel reduction, because unstressed syllables are articulated with less effort. However, Lindblom argues that unstressed syllables are reduced because they are articulated in less time and not due to less articulatory effort. He substantiates this with the results of a small follow-up investigation, where one speaker had to utter stressed syllables at speech rates from 0.5 to 6 syl/s. Even though the syllables are stressed, target undershoot occurs to the same extent as for unstressed syllables. Lindblom concludes that there are limitations inherent to the articulation system which are independent of articulatory effort and are based solely on duration.

This point of view, according to which duration is the main factor for vowel reduction, is not unanimously shared in the literature. In [Del69] the author argues that the main deter-

minants of vowel reduction are stress and speech rate rather than duration because a shorter duration is only a product of less stress and a higher speech rate. Therefore, duration should be seen only as a secondary determinant of vowel reduction [Del69, p. 298].

In experiments reported by Fourakis [Fou91] this argumentation could not be sustained. He found that stress and duration had only a marginal effect on vowel reduction while context seemed to be the most important factor in determining the position of the vowel in the vowel space. Similar results were obtained by Van Bergem [vB93a], who observed a strong influence of the context on reduction apart from stress and word class. For example, vowels occurring within the articulatorily neutral context [hVd] are farther away from the vowel space centre than vowels with less neutral context and their distance was found to be unaffected by duration. Van Bergem therefore argues that reduced stress shifts the formant frequencies in the direction of their neighbouring segments rather than towards the centre of the acoustic vowel space. If the context is neutral, no reduction will take place.

More recently, another factor that might have an influence on vowel reduction was investigated by Aylett [Ayl00]. It is based on the assumption that the overall redundancy of spoken speech is smoothly distributed. This means that language redundancy and care of articulation are inversely related. According to this idea high-frequency words that do not convey new or important information, such as articles or prepositions, are produced with strongly reduced formant frequencies. In contrast, words that cannot be predicted from the context are articulated carefully. Aylett analysed spontaneous speech from an English task-oriented dialogue corpus and computed two measures which are inversely related to reduction which he calls care of articulation.

The first one measures the Euclidean distance from a vowel token to the speaker specific vowel space centre in a mean and variance normalised, bark scaled frequency space. Thus, the further away a vowel token is from the vowel space centre, the less reduced it is. The mean and variance normalisation was performed in order to adjust for different speakers' vowel spaces in terms of absolute formant frequencies, but also in terms of the frequency ranges that the formants can take. Since this measure does not take into account any coarticulatory effects, another measure is introduced, which measures the relative distance of a vowel token from the quasi target frequencies of the vowels. These target frequencies are obtained from words spoken clearly and in isolation by the test speakers. The further away a vowel token is from these target frequencies, the more coarticulated it is.

In the final analysis of the relationship between redundancy and care of articulation Aylett showed that the reduction measures were indeed correlated with the measures for language redundancy. Thus, low frequency words or words that are not predictable from the preceding words have formant values that are closer to those of the clearly spoken vowels than high frequency words or words that can easily be predicted from the context. In a discussion of the relationship of care of articulation and duration the author states, that both phenomena

are closely related. However, he also argues, that duration and care of articulation are not the same and do not necessarily have to behave in the same way.

While the stress pattern and the phonetic context are rather linguistic variables that influence the acoustic characteristics of vowels, there are also a variety of non-linguistic factors that affect the realisation of spontaneous speech. For example, complex interactions with phonetic reduction have been reported for different kinds of emotion. It is generally assumed that changes in the emotional state of a speaker affect his or her physiology and consequently the acoustic characteristics of his or her speech [GTS88].

In an investigation of the effects of different emotions on articulatory reduction in terms of segment deletions and assimilations it was shown that the emotional states of fear and sadness were both correlated with high articulatory reduction [KPS99]. However, while fear was characterised by a fast speech rate, which is consistent with the finding of reduction, sadness showed a slow speech rate together with reduction. This finding is in contrast to the previously reported results, where reduction did only occur in fast or unstressed speech. While in this analysis reduction is measured on a symbolic level by deletions and phoneme alterations, there are indications that they are accompanied by acoustic reduction. Although no articulatory measurements were taken, the emotional state of sadness was interpreted to correlate with low muscular tension, which causes a high amount of assimilation and deletions. If one follows this argumentation, acoustical reduction in terms of reduced formant frequencies would be expected to coincide with deletions and assimilations, because low muscular tension of the articulators is likely to cause articulatory undershoot. In this case a low speaking rate as observed for the sad emotional state would be accompanied by reduced formant frequencies. This is contrary to results from former investigations, where reduction is correlated with a high speaking rate or shorter segmental durations, and indicates that other factors than duration or stress can cause reduction, too. However, it should be noted that these analyses were carried out on speech produced by actors. Although the analysed speech proved to be identified as fearful or sad in a listening experiment, it remains to be shown that natural emotions by naive speakers evoke the same acoustical characteristics of the speech.

Also, apparently similar emotional states such as different kinds of psychological stress do not have similar effects. Cognitive stress induced e.g. by a time-critical task to be performed while speaking leads to an increase of the precision of articulation. In contrast, emotional stress induced by pictures of severe accident injuries is correlated with acoustical reduction [TS86]. Thus, the relation between the emotional state of the speaker and acoustic reduction is a complex one with different emotional states evoking different acoustic effects.

In summary, it can be stated that there are different causes for acoustical reduction. While duration or articulation rate is one of the most obvious causes, other factors such as linguistic stress, speaking style and even the emotional state of the speaker influence the articulatory

and acoustic precision of spontaneous speech. All these findings show that reduction is a very common phenomenon in many different kinds of speech. Therefore, in the following sections a closer look will be taken at the acoustic characteristics that are associated with reduction.

3.2.2. Centralisation

In most of the above mentioned investigations reduction was measured by means of the relative distance of a vowel token to the vowel space centre. This implicitly suggests that each token is shifted towards this centre. However, as has already been pointed out, reduction is actually caused by increased coarticulation of neighbouring phonemes. This would mean that the formant frequencies rather shift towards the values of their neighbours instead of the neutral vowel in the centre of the vowel space. In the case of such a centralising shift, the vowel space centre would play an important role in reduction and in predicting the places of the vowels in the acoustic vowel space. In the literature some discussions about the role of the vowel space centre in reduction can be found. Some of the arguments will be briefly summarised in this section.

A very radical view concerning the role of the vowel space centre, as represented by the neutral vowel [@], is taken by Van Bergem in [vB95]. He argues that there exists no neutral vowel in the centre of the vowel space. He states that the results of former experiments where a shift of the formant frequencies of the vowels towards the vowel space centre was observed were artifacts caused by the summation over all vowel tokens regardless of their context. In his experiments he found that a vowel may be completely assimilated with its phonemic context while its vowel colour can still vary widely. In order to show this effect, Van Bergem measured the curvature of the formant tracks of vowels over time. He observed that in fast speech there was an especially strong effect on the curvature of the neutral vowel [@] where the formant tracks of the vowel approached a straight line between the on- and offsets. This can be interpreted as the vowel having no target of its own. From this Van Bergem draws the conclusion that the vowel [@] is not produced with a neutral vocal tract, but that it is rather a vowel that is completely assimilated with its phonemic context. Therefore, he argues that vowel reduction is not a centralisation tendency where all vowels tend to become the same neutral vowel produced with a neutral vocal tract, but that it is rather an increase in contextual assimilation. According to this argumentation the neutral vowel is the most convenient point in the articulatory-acoustical space to go from one consonant to another and depends therefore only on the context of the vowel. Therefore, it can be argued that other reduced vowels are not shifted towards the centre, because there exists no neutral vowel [@]. They are rather shifted towards the phonemic context in a similar way as the neutral vowel is. The position of a strongly reduced vowel will therefore only depend upon its context.

However, this point of view is not undisputed in the literature. In a study, where the focus was put on the target of [@] as expected to be observable in slow speech, Barry investigated the influence of the consonantal and vocalic context on the formant frequencies of the [@] [Bar98]. As in the former study, the results showed that in fast speech the position of the [@] in the F1-F2 plane is shifted towards its consonantal and vocalic context. While both investigations achieved the same results the former fails to explain why there exists a stronger curvature in the formant tracks of slow [@]. For slow speech this means that the formant values of [@] are less affected by their context and constitute a target of their own. For the question of the role of centralisation in fast speech this only indicates that [@] is also affected by its phonetic context. However, in contrast to other vowels, the shift of [@] will point rather into the direction of the outer boundaries of the vowel space.

In [KVB80] the problem of the role of [@] is avoided by making use of the vowel space centre which is computed as the mean over all vowels. In her data, which consisted of speech ranging from conversational speech to words spoken in isolation, Koopmans-Van Beinum found a constant relationship of the first and second formants of this centroid which represents the neutral position of the articulatory organs and is therefore anatomically determined by the speaker's vocal tract. This centroid was used to measure the amount of reduction by computing the total variance of all vowels in different speech conditions which gives the acoustic system contrast. This measurement implicitly assumes that reduction is based on a shift of the vowel formant frequencies towards the centre.

The analysis was carried out on Dutch vowels spoken in various speech conditions ranging from vowels spoken in isolation to normal conversation. The results indicate that the acoustic system contrast is most reduced in unstressed vowels in free conversation while least reduction can be found in vowels produced in isolation. Here, centralisation means less articulatory effort which finally leads to vowels produced with a neutral vocal tract. This means that strongly reduced vowels contain more information about the speaker's vocal tract and less information about the vowel identity.

However, this interpretation of the centralisation point to represent the speaker-specific neutral vocal tract position is put into perspective by an experiment where the role of the centroid in different languages was investigated [Del69]. From the comparison of acoustic and articulatory data of the four languages English, German, Spanish, and French, Delattre concluded that the formant frequencies of unstressed vowels shift towards a language specific pole near the centre of the vowel triangle. Thus, the speaker-specific centralisation point as proposed by Koopmans-Van Beinum seems to be at least influenced by language-specific characteristics.

In summary, while most investigations are based on the implicit assumption that acoustic reduction is caused by an increased coarticulation, the measurement of reduction is generally based on the distance of the vowel tokens or the mean formant frequencies to the vowel space

centre. This indicates that apart from an increased coarticulatory effect, reduction means also a shift towards the acoustic vowel space centre. However, even studies where the difference between an increase in coarticulation and centralisation is discussed fail to analyse in detail the relative amount of coarticulation versus centralisation to the overall effect of reduction.

3.2.3. Effects on Dynamic Features

In the experiments reported so far the analysed formant frequencies were taken from the middle of the vowel segments. However, articulation is a continuous process and there is evidence that human listeners are able to recognise a vowel even before the target frequencies of this vowel are reached. This indicates that there is more information in a vowel than may be derived from the static properties derived from the middle of the segment. For example, more information can be drawn from the change of the formant frequencies over time. In order to capture such time varying information of formants, several investigations focus on dynamic formant movements which are captured either by a mathematical model of the formant tracks or by analysing the formant frequencies at different time positions in the vowels. From these dynamic characteristics it is possible to draw conclusions about the movements of the articulators.

In an attempt to explicitly capture the dynamics of the formant movements Van Bergem [vB93a] modelled the formant tracks over time with second order polynomials. Second order polynomials are able to model tracks with one minimum or maximum. The underlying assumption is that the formant track in a vowel starts from a neutral point, which might be influenced by the context, reaches a maximum in the middle of the segment, and will finally go back to a neutral point. He compared the formant tracks of vowels spoken in isolation with vowels where the steady state part was heavily reduced. In order to obtain comparable results, a time normalisation of the formant tracks was performed by scaling all movements to fit the same time window[1].

The computation of the second order polynomials showed that the time normalised tracks were not identical. The normalised formant tracks were flatter for reduced vowels as compared to un-reduced vowels. This was interpreted to show that reduced vowels are articulated with even slower movements than un-reduced vowels. However, since the method of the time normalisation as a crucial point in this investigation remains unclear, this result may only be an artifact.

[1] However, note that it remains unclear whether this time normalisation only means a change of the temporal resolution or includes the frequency range as well. If only the time resolution is changed, the normalisation would severely affect the curvature of the formant tracks. In this case, a longer duration would lead to a steeper movement of the formants. The time normalised formant tracks are only comparable, if the time normalisation included both the temporal and the frequency dimension.

All the above mentioned investigations were able to find some kind of reduction that occurred in fast or unstressed speech. However, the conditions under which the recordings of the analysed speech samples took place are rarely considered. In a revised version of the target undershoot model an additional factor was, therefore, taken into account: the speaking style. Moon & Lindblom [ML94] showed that speakers are able to talk faster without necessarily reducing the movements of their articulators. In order to obtain speech data that was likely to produce such an effect, they asked their test speakers to talk fast and very clearly. The results showed that, while the test speakers did indeed reduce the duration of their segments in order to speak faster, they did not reduce the formant frequencies of their vowels. They achieved this by faster movements of the articulators, which is reflected in steeper slopes of the formant tracks.

Similar results were obtained by Pols & Van Son [PvS93], who compared the speech of a professional news reader reading at normal and fast rate. They found no indication of undershoot caused by shorter segment duration. Instead, the formant tracks were shown to be steeper in the fast speaking condition. Therefore, the authors proposed an *active model* of speech production as opposed to the target undershoot model (see Fig. 3.3). The active model is characterised by a more heavily curved formant movement for fast speech, while the target undershoot model predicts similar slopes for the formant tracks for both rate conditions.

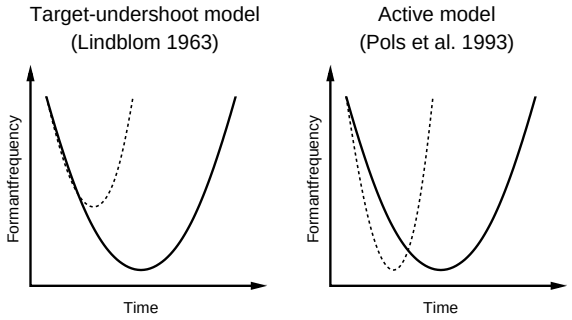

Figure 3.3.: Active Model as proposed by Pols & Van Son[PvS93] in comparison to Lindblom's target undershoot model. The target-undershoot model predicts a reduced formant frequency with the formant movements remaining stable for fast and slow speech. In contrast, the active model predicts similar formant frequencies for fast speech which are achieved by steeper movements.

An example of stable formant movements for slow and fast speech is given in an analysis of the on- and offset formant frequencies of American English diphthongs, where it was found that the change of the second formant remains fixed in faster speech together with a stable onset frequency [Gay68]. A compensation of the reduction of available time was achieved by a change of the offset formant frequencies. Similar results were obtained

in a follow-up study of monophthongs [Gay78]. It was shown that the midpoint formant frequencies as well as the formant movements were not affected by speech rate. However, this was achieved by the onset formant frequencies lying closer to the target frequencies. This means that there are different reduction strategies for diphthongs and monophthongs. While in diphthongs the offset frequency is shifted towards the target frequency, in monophthongs this compensation seems to take place at the onset of the vowel. These results are difficult to interpret, since they indicate that the compensation of the reduced time is shifted towards the beginning or end of the vowels, which means that the adjacent consonants have to be produced faster or reduced or have to incorporate more influences of the vowel. This indicates that consonants are also heavily affected by reduction.

In summary, the results of the investigations of the formant movements over time show that in the case where a faster speaking rate induces articulatory and acoustic reduction, the formant movements tend to remain stable, indicating a similar velocity of the articulatory movements for fast and slow speech. In fast speech the shorter duration is compensated by an earlier cut off of the articulation before the target position is reached, which results in reduced formant frequencies. However, it has also been shown that this strategy can be changed by a different speaking style. If the speaker attempts to speak clearly and fast the target values can be reached even in a shorter period of time by faster movements of the articulators, which is reflected in steeper formant tracks. It has finally been shown, that the consonantal influence on the acoustic characteristics of the vowels in fast speech is not entirely clear. There are some indications that reduction also takes place in consonants. This topic will be addressed in the following section.

3.2.4. Consonant Reduction

Since reduction is mostly defined and analysed in vowels, the question arises what happens with the consonants in fast or unstressed speech. The observed effects of articulation rate on vowels mostly concern the formant frequencies. It is therefore necessary to find new acoustic features that reflect articulatory reduction in consonants. The definition of new measures for such effects will, therefore, receive a lot of attention in the following section.

Consonant reduction was investigated in a series of experiments by Van Son & Pols [vSP95] [vSP96] [vSP99], who defined several parameters that they assumed likely to be affected by consonant reduction. Their analysis is based on a corpus of spontaneous and read speech of a professional news speaker which contains a read version of spontaneously produced text. Thus, the two compared speaking styles consisted of a version which was assumed to contain heavily reduced segments and a read version as reference with less reduction.

In order to measure the coarticulatory strength, which is assumed to be higher in reduced

phonemes, they calculated the differences of the slopes of the second formant (F2SD) at the end and the beginning of a consonant in a VCV sequence (cf. Fig. 3.4).

$$F2SD = \Delta F2_{End} - \Delta F2_{Beg} \tag{3.1}$$

If the F2 values of the consonant target lie between those of the adjacent vowels, coarticulation is assumed to be high and the F2SD values tend towards zero (Fig. 3.5 right). If the F2 values of the consonant constitute a minimum or a maximum, the articulation is assumed to be carried out more carefully and with less coarticulation because the consonant represents its own target. Even if an unvoiced consonant is articulated, it is assumed that the formant movement at the end of the preceding vowel indicates the movement of the articulators towards the targets for the consonant. This is characterised by higher values of the slope difference (Fig. 3.5 left). It was shown that the F2SD values were indeed lower for spontaneous speech especially for plosives and fricatives. This indicates that in spontaneous speech the frequencies of the second formant of consonants tend to lie between those of the surrounding vowels while in read speech they constitute their own targets.

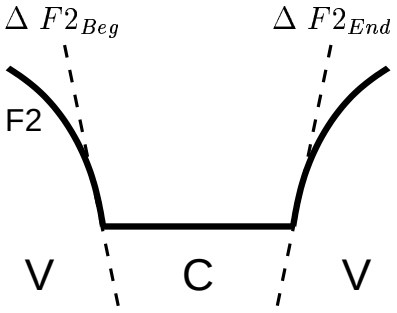

Figure 3.4.: F2 slope difference to measure the extent of coarticulation in consonants [vSP99].

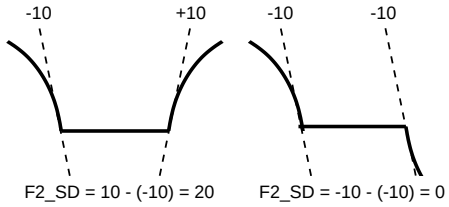

Figure 3.5.: Examples of F2 slope differences. The left example shows an accurate articulation with no coarticulation, the right example shows a highly coarticulated consonantal segment, with the formant frequencies lying between those of the neighbouring vowels.

Another measure for coarticulation are the F2 locus equations based on the notion of Lindblom's target-locus distance. The F2 locus equations describe the correlation between the F2 value at the vowel onset and the observed vowel target. In this investigation the target is defined as the formant frequency where the most extreme values of the realisation is achieved [vSP96]. It is therefore not to be confused with the target definition by Lindblom [Lin63]. While vowel onset and target will be strongly correlated in all situations it is assumed that the variance will be higher in spontaneous speech. However, the results showed that the overall differences in correlation between spontaneous and read speech were small

and not consistent. An analysis of the influence of duration on this measure was not carried out.

A spectrum-based measure proved to be a better indicator of consonant reduction. The centre of gravity (COG) of the power spectrum is strongly correlated with perceived syllable stress [SVH96]. It is the average frequency weighted by the acoustic power:

$$COG = \frac{\sum_i f_i \cdot E_i}{\sum_i E_i} \tag{3.2}$$

where f_i is the centre frequency in Hz of the i-th frequency channel and E_i the spectral power as a function of the frequency (cf. Fig. 3.6). The COG is a good indicator of the steepness of the spectral tilt which is closely correlated with vocal effort. According to Van Son the spectral slope is determined by the steepness of the glottal flow velocity change at the closure of the vocal folds. Thus, the steeper the harmonic source spectrum produced by the glottis, the steeper the filtered spectrum at the end of the vocal tract will be. A lower COG indicates a steeper spectrum and is therefore interpreted to be caused by less effort and consequently expected to indicate reduction. This was indeed what the results showed: the mean COG of the different consonants were all lower in spontaneous speech as compared to those of read speech.

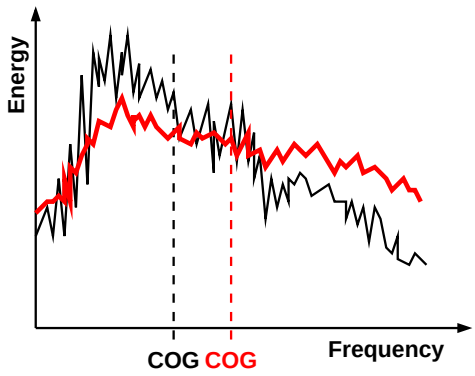

Figure 3.6.: Centre of gravity of the power spectrum. The black line shows the spectrum of a reduced vowel or consonant with a steeper power-spectrum which results in a lower COG. The red line shows a more level power-spectrum which is characteristic for clearly articulated vowels or consonants.

Furthermore, the intervocalic sound energy difference was analysed [vSP96]. This measure was defined as the difference between energy of a consonant and the mean energy of the surrounding vowels. It was argued that vowel reduction correlates with reduced intensity. If this reduction only affected vowels but not consonants, the differences between the

energy levels of vowels and consonants should be diminished. However, although there was a measurable effect on the intervocalic sound energy difference, this effect was too small to be reliable. It was therefore assumed, that the energy of consonants is reduced by almost the same amount as vowels in spontaneous speech.

In summary, these results show that consonants are also affected by reduction due to changes of the speaking style and in duration. However, these effects are only measurable by more indirect approaches. For example, the measurement of the F2 movements of the surrounding vowels can indicate consonantal reduction based on the assumption that they reflect the articulatory movements towards the consonantal targets. Apart from such formant-based measures the acoustic correlate of speech effort as captured by the mean spectral frequency seems to be a good indicator of reduction. However, for a measure based on the change of energy over time in order to measure the energy differences between vowels and consonants no reliable results were obtained.

3.3. Perceptual Effects of Speaking Rate

As has been shown both the durational and the spectral characteristics of the speech signal are heavily affected by changes in the articulation rate. This raises the question as to how human listeners compensate for these effects. In this section some evidences for the existence of such a compensation mechanism will be briefly presented. These pieces of evidence support different kinds of theories about the underlying mechanisms of perception. There are mainly two opposing theories concerned with how timing information is incorporated during the decoding of the speech signal. Intrinsic timing theories assume that there exist features that are independent of speaking rate. Evidence for such features are effects found to occur in durational normalisation. In contrast, extrinsic timing theories are based on the assumption, that before spectral normalisation takes place a rate detection is applied which indicates how the spectral normalisation has to be carried out. These theories are based on the observation that human listeners are very good at normalising for spectral degradations due to speaking rate variations. In [WMV94] an integration of both theories is suggested by the assumption that intrinsic timing occurs at syllable level while extrinsic timing mechanisms are applied at a more global level because it requires a rather long term rate detection.

3.3.1. Durational Normalisation

Since many phonemic contrasts are based on durational differences, e.g. different voice onset times for voiced and voiceless plosives, this leads to the question how listeners normalise for articulation rate changes in conversational speech. Perceptual experiments show that listeners

process durational cues in a rate-dependent manner.

For example, one important distinguishing characteristic of the syllable-initial bilabial consonants [b] and [w] is the abruptness of the consonantal onset which is reflected in the initial formant transitions. Syllables with short initial transitions are perceived as beginning with a stop consonant whereas those with longer transitions are perceived as beginning with a semi-vowel. In [ML79] synthesised stimuli with different durations of the formant transitions were presented to listeners who had to decide whether they heard a [b] or a [w]. Apart from the transition phase the overall duration of the syllable was varied, too. The classification results of the listeners show that with an increasing duration of the syllable a longer transition phase is required to perceive [wa] as opposed to [ba].

Similar results were obtained for the discrimination of the syllable-initial bilabial plosives [b] and [p] [PD82]. One relevant acoustical cue for the discrimination of voiced and voiceless plosives is the voice onset time (VOT). The VOT measures the time between the release of the closure and the beginning of the voicing for the following vowel. The VOT for voiced obstruents was found to be shorter than for voiceless ones. It was shown that the ratios of the VOT of voiced plosives to syllable duration are lower than those of voiceless plosives across different speech rates. These ratios proved to be relatively constant for each, voiceless and voiced consonants, over different speech rates. The authors therefore conclude that such VOT-syllable duration ratios are consonant specific constants which are independent of the articulation rate.

3.3.2. Spectral Normalisation

The rather severe effects of spectral reduction on the spectral characteristics of both vowels and consonants lead to the question if and how listeners normalise for these acoustic variations.

Although listeners are able to correctly identify reduced vowels in fast speech they cannot do so without the surrounding context which carries further rate information [VSSE76]. When a syllable containing a tense vowel is taken from a fast speech sample and presented in isolation, the vowel is likely to be identified as the lax version of the vowel although the vowel token is perceived as tense when surrounded by fast speech. The perception of a lax vowel is even stronger when the fast syllable is presented in a slow carrier sentence. This indicates that listeners use rate information when judging vowel quality. However, when a syllable from slow speech is presented in a fast carrier sentence the recognition performance does not decrease [VS77]. Slow syllables tend to be correctly identified regardless of their surrounding context. This is explained with the assumption that a slow syllable contains enough information about its speech rate and can therefore not be perceived as fast. These results suggest that human perception involves several mechanisms to compensate for vowel

and consonant reduction in conversational speech. They are examples for extrinsic timing theories because they suggest that the articulation rate has to be known in order to correctly identify a vowel as lax or tense.

3.4. Summary

In this chapter it has been shown that there exist a variety of effects that speech rate variation has on the acoustic-phonetic properties of speech. Analyses of the temporal properties of phonetic segments in spontaneous speech show that variation of the overall speech rate is reflected in both the number and length of pauses that are inserted between words and the articulation rate, which denotes the number of units uttered per second. Although closely correlated to the articulation rate, the articulation speed denotes a further variable in speech rate variations. Thus, a fast speech rate is achieved by making less and shorter pauses but also by shortening the phones' durations. This does not necessarily mean that the articulators have to move faster. A compensation of less time available for a certain number of speech segments can be achieved by a less accurate articulation where the target positions, and therefore the target frequencies of the formants, are not reached.

Even more importantly, these temporal changes, which co-occur with speech rate variations, cause severe spectral degradations known as reduction. Reduction is generally described as a shift of the vowel formant frequencies towards the vowel space centre. It is also commonly agreed upon that reduction is caused by increased coarticulation in fast speech, which means that the formant frequencies are shifted towards those of the phonetic context. However, in the literature no attempt has been reported to measure the relative contribution of both effects, centralisation and coarticulation, on reduction.

Although reduction is strongly correlated with segment duration and articulation rate, there exist several different factors which can cause spectral degradations as well, such as the lack of stress, semantic content or certain emotional states of the speaker. On the other hand, a faster articulation rate is not necessarily accompanied by reduction. If the speaker is aiming at a clear speaking style he or she can compensate for the shorter segment duration by faster movements of the articulators. This is acoustically reflected in steeper formant movements. Thus, there exists no one-to-one relationship between duration and reduction. It is rather the case that both measures together, duration and the amount of reduction, are good predictors of the actual acoustic realisations.

Apart from acoustic reduction of vowels there exists evidence for the acoustic reduction of consonants, too. Since formants are not a suitable measure for consonants, the measurement of consonantal reduction is based on different features. One of the most effective measures of consonantal reduction is based on the effects of speech effort. Variations of speech

effort are reflected in the tilt of the power spectrum. However, by a more indirect approach where the formant movements of the surrounding vowels are taken into account, even a formant based measure is able to indicate consonant reduction.

In summary, speech rate variation has been shown to affect not only durational but also spectral features of the speech signal. These effects are heavily correlated with duration but also with different other factors. Apart from vowel reduction, spectral degradations can be observed and measured with different measures in consonants, too. While these results strongly suggest a systematic relationship between speech rate variation and diverse acoustic correlates, the following chapter will show that these systematic relationships are generally not taken into account in the modelling of speech rate variation in automatic speech recognition.

4. Speech Rate Modelling in Automatic Speech Recognition

In automatic speech recognition the severe performance degradations caused by speech rate variations have motivated many attempts to model speech rate. They can be divided into modelling strategies on the symbolic level and those focusing on the acoustic level. Since a fast speech rate tends to cause deletions and coarticulatory effects that can be captured on the symbolic level, some approaches focus on an explicit modelling strategy by using different variants of pronunciation. In contrast, acoustic changes that cannot be captured on the symbolic level, such as reduction and coarticulation, are modelled by directly adapting the acoustic models to the different speech rates. Since this work concentrates on the acoustic effects of reduction and coarticulation a focus will be laid on the latter approaches.

All modelling schemes that focus on modelling the spectral effects of rate variation make use of some kind of speech rate estimation upon which the training of rate dependent models or the choice of rate adaptation schemes for recognition are based. In the literature a variety of speech rate measures can be found of which the most commonly used measures will be described in the following section. The different techniques for adapting models or features to the measured speech rate are presented in the second section.

4.1. Speech-rate measures

In general the rate measures can be divided into two groups. The most obvious approach is to measure rate in terms of phone or syllable rate. However, such approaches need to perform a preliminary segmentation of the test utterance. Since for this first recognition pass no rate specific models are available the segmentation is performed by the rate independent models. Therefore, such measures are error prone. Attempts have been made to overcome these shortcomings of a preliminary segmentation with the definition of several peak counting measures which are based directly on the signal and do not need a prior segmentation pass. Other approaches avoid the segmentation problem by measuring rather the effects of speech rate variation than rate itself. Several kinds of rate measures are based on the spectral or acoustic features. Such measures rely on the observation that speech rate not only affects

the time domain of the speech signal but also the spectral characteristics.

In this section an overview of some rate measures will be given. A start will be made with measures of rate in terms of phone and syllable rate. Finally, some less usual measures based on the changes in the spectral characteristics will be presented.

Phone rate. A common measure for the estimation of speech rate is the phone rate, which is defined as the number of phones per second. Since the intrinsic duration can vary widely between the different phones of a language, some approaches perform a normalisation by dividing the observed phone duration by the average duration of this phone [RHAH99]. This procedure implies, that the utterance has to be recognised completely before rate dependent recognition can be performed. Furthermore, it is vitally important to have a good accuracy of the phone identification, since a normalisation by the wrong factor can have severe effects on the rate estimation and therefore on the further processing. Therefore, most approaches make use of an unnormalised phone rate by detecting phone boundaries and counting the segments without assigning the phone identity [MTÁ98] [MFM95] [VM96]. These approaches yield classification accuracies of 70% to 80% in [MTÁ98] and of 73% in [MFM95]. It was also reported that the classification was best for diphthongs and glides and most difficult for voiceless consonants [MFM95]. The best classification result was achieved for speech with an average rate, while fast speech was most difficult to classify.

Enrate. A more sophisticated measure which does not require a preliminary segmentation of the test utterance is the energy rate, or enrate, which is assumed to be closely correlated with the syllable rate [MFM97]. It is basically the spectral moment of the speech signal low-pass filtered at 16 Hz. In other words the enrate denotes the weighted mean frequency of the energy envelope variations over a time window of one or two seconds. These energy variations roughly coincide with the syllable nuclei. Their frequency is therefore assumed to represent the syllable rate. It has been shown that the enrate is only moderately correlated with the syllable rate but is a good predictor for increased recognition errors which makes it well suited for speech rate adaptation in speech recognition. An enhancement of this measure is achieved by the addition of two peak counting measures [MFL98]. The first one is a count of the energy peaks from the whole frequency range of the spectrum while the second one is based on the counting of energy peaks in the different frequency sub-bands. High energy peaks in all frequency bands lead to a high value and indicate the existence of a vowel. The count of these peaks should therefore be strongly correlated to the syllable rate. The enhanced measure, the *mrate*, is defined as the average value of the enrate and both peak counting measures. It shows a correlation of .67 with the syllable rate. A classifier based on the mrate was reported to achieve an accuracy of 58% in a decision task where the utterance had to be classified as fast, average or slow. This measure has the advantage of avoiding a

preliminary segmentation of the test utterance. It is furthermore based on the syllable rate which represents a larger phonetic unit than the phone rate.

Modified Loudness. Another approach to estimating the syllable rate is taken in [PR98], where the rate estimation is based on the observation that the energy of vowels is distributed mainly in the lower frequency bands while consonants exhibit more energy in the higher frequencies. Thus, a high positive difference between the energy in the low and high frequency bands indicates vowels or syllable nuclei. The modified loudness is defined as the difference of the energy in the low and high frequency bands [WR89]. The peaks of this difference therefore denote vowels or syllable nuclei, so that a peak counting of the modified loudness is supposed to be correlated with the syllable rate. It has been shown that this measure has a correlation of .79 with the syllable rate as derived from a manual segmentation of the speech signal on the Verbmobil Corpus [KLP+94]. Therefore, the peak counts of the modified loudness seem to be a good predictor of the speech rate. Along with the other estimates of the syllable or phone rate this measure does not take into account whether deletions which are likely to appear in fast speech have taken place. It only measures the rate at which syllables actually occur. Since these measures are generally only applied on parts of speech without silences they are in fact estimates of the articulation rate.

Dynamic features. Apart from measures of the rate of certain phonetic units the speech rate can also be estimated by the effects that rate variations have on the acoustic features of the speech signal. In [MTÁ98] it has been shown that it is possible to classify speech as fast or slow by training a classifier with the speech samples, that were previously rated as fast, average or slow by a phone rate measure. Experiments showed, that the dynamic cepstral features are most affected by speech rate variations while the static MFCCs did not show a systematic effect[1]. Therefore, only the dynamic features were used to train the classifier. By this procedure the classification of utterances as slow, average or fast yielded an accuracy of about 80% for slow speech and 70% for fast speech. The classification was counted as correct when it was classified into the same class as by a phone rate measure. Thus, more than two thirds of the utterances can be correctly classified into phone rate classes solely by their dynamic features.

[1] However, it should be noted that these results were obtained on a corpus specially designed for rate analyses. It consists of utterances which were meant to be either fast, normal or slow. It is possible that this data gathering method evoked a clear speaking style where no undershoot occurs due to faster movements of the articulators. As pointed out in the previous chapter, this affects especially the dynamic features at least in terms of formant movements.

Distance between successive feature vectors. A similar approach is taken in [TY99] where another feature based rate measure is proposed. The underlying assumption is that fast speech rate is accompanied by a fast change of the feature vectors. In the experiment the dependency of neighbouring vectors is measured by computing the Euclidean distance between successive frames in the feature space. The feature vectors are composed of static features plus the first and second order derivatives and an overall energy component. Thus, successive feature vectors are supposed to be less dependent in fast speech as compared to slow speech. A comparison with the normalised phone duration revealed that both measures were linearly correlated: the shorter the phone duration is, the higher are the Euclidean distances between the feature vectors. However, a classifier based on this distance was only able to classify 57% of the frames or 60% of the phones correctly as fast or slow which is only little above chance level. By separating the training data according to the distance measure instead of the phone duration the classification was reported to be significantly enhanced. However, results were only reported for one diphthong. These results indicate, that while speech rate and acoustic degradations are closely related, a direct measurement of the acoustic effects yields a better classification of slow and fast speech when the classification is based on the acoustic features.

It is interesting to note that a seemingly opposed prediction of the correlation of successive feature vectors did also hold true in another investigation. In [KB99] it was argued that in fast speech the amount of coarticulation increases which should be reflected in the predictability of the information conveyed by neighbouring feature vectors. In order to measure the amount of coarticulation, the mutual information of adjacent feature vectors consisting of static and dynamic MFCCs was computed in different speech conditions such as fast and slow . The mutual information is an information theoretical measure that gives an indication of the extent to how predictable a signal Y is, given knowledge of a signal X. The higher the mutual information between two signals is, the better predictable they are from each other. In accordance with this assumption, it was shown that the mutual information of neighbouring speech frames is higher in fast speech which indicates that the feature vectors are more predictable and therefore more correlated in fast speech.

Given these seemingly opposed results the question arises whether they are due to different corpora. It might be that the first corpus did not show an increased coarticulation with a higher speech rate while the second corpus did. This would explain the different results since the first measure aims at capturing the speech rate while the conditional mutual information measures coarticulation. However, it is more likely that the two results are not as opposed as they appear, since the Euclidean distance is not necessarily related with the predictability of two vectors. Thus, while in fast speech the feature vectors are further away from each other in the feature space, they still convey more information about the next one. This is because both measures focus on different aspects of articulation. The distance of the features repre-

sents the articulator speed. The faster the articulation is, the faster will the movement of the features in the feature space be. On the other hand the mutual information is a measure for coarticulation which is not identical with the rate of articulation. A higher extent of coarticulation means that some articulatory and acoustic features remain over a longer stretch of speech than the phone for which they are characteristic. This indicates that speech rate and coarticulation have different effects on the acoustic features of the speech signal.

This shows that the definition of the phenomenon that is intended to be captured is crucial. A change in the definition leads to severe performance losses. Thus, either the phone rate is the base for the rate measurement, then the measure should rely directly on the duration of the phones. Or the articulator speed is measured by the distance of the feature vectors. In this case, a different phenomenon is taken into account, since phone rate and articulator speed are two different measures.

Applicational aspects of rate estimation. In most of the experiments presented in the literature the speech rate was computed as a mean over the whole utterance [MFM95] [VM96] [MTÁ98]. This assumes that speech rate is a speaker specific variable which remains relatively stable over an utterance. However, this assumption can be problematic since the local speech rate variation is significant especially in spontaneous speech. A running estimate of the speech rate therefore provides a more detailed representation of the rate variation.

However, for running rate estimations the window length can cause severe problems for the first and last frames of an utterance. For example, for the enrate a low-pass filtering of 16 Hz is performed by applying a time window of around one second in order to capture frequencies of down to 4 or 5 Hz which represent very slow syllable rates [MFM97]. While rate changes in the middle of the sentence can be captured well with these measures, variations at the beginning and end of a sentence remain unnoticed because of the lack of context. The values for these frames are generally interpolated by the first and last available rate estimates of the utterance. Thus, while it would be desirable to perform a running estimate of the speech rate in order to capture local variations, this approach is generally not followed for practical reasons.

4.2. Compensation Techniques

There are two ways in which adaptation can be performed. The most common approach consists in changing the model parameters according to the speech rate. For example, if a rate estimation prior to decoding detects a fast speech rate, the model parameter can be adjusted in order to model the acoustic effects. A very rarely applied method is the modification of the observed feature vector according to the speech rate. In this case the features have to be

modified in order to fit the trained models better.

4.2.1. Model Adaptation

Approaches to rate modelling generally rely on the principle of *rate-dependent models* . A speech rate estimation is performed on the training set which is then divided into three or four subsets which represent the different rate classes. These subsets are used to either train completely independent sets of parameters or to adapt general models to specific speech rates. Similarly, recognition is performed in two phases. First, a rate estimation is performed on the speech signal. This can already be a time critical factor, since some approaches need to analyse the whole utterance in order to determine a mean rate over the entire utterance. Upon the completion of the rate estimation the appropriate parameter set which has been trained with data of the same speech rate is chosen. Only then is decoding initiated with speech rate dependent models. However, there exist a variety of different model parameters which are affected by speech rate variation and are therefore subject to modification. In some approaches, all parameters of the whole acoustic model are adapted, while other approaches focus on the features affected most.

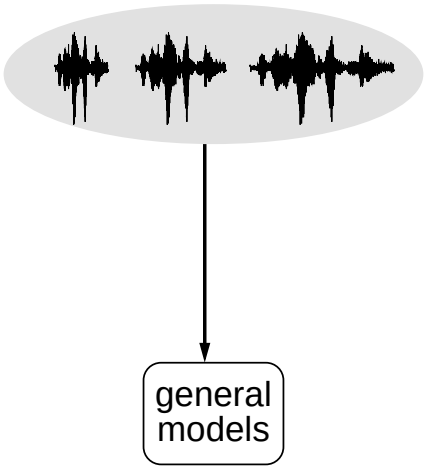

Figure 4.1.: Training of general models with the whole training data available.

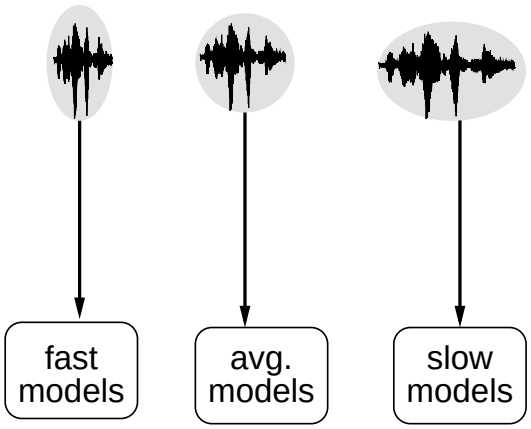

Figure 4.2.: Training of rate dependent models.

As shown in Figure 4.2, the approach of rate-dependent models relies on a few discrete rate classes only. This makes it possible to train different sets of parameters without having to explicitly extract the transformation that the features undergo with a change of the speech rate. Almost all adaptation schemes reported in the literature make use of discrete rate classes. However, as can also be seen from Figure 4.1 and 4.2 the separation of the training material severely reduces the available training data to a fraction of the original

amount. The more rate classes are trained, the less training data will be available for the rate dependent models. Therefore, some approaches focus on handling sparse data problems.

The estimation of speech rate in the decoding phase is another crucial factor in speech rate modelling. As can be seen in Figure 4.3 the rate estimation determines which model is chosen for decoding. If this estimation is error-prone, the wrong models are likely to be selected which can degrade the performance severely. The more rate classes are trained, the more difficult it will become to choose the correct rate class. However, generally not more than two or three rate classes are used, because of the dramatic decrease in the amount of training data.

In the following section some approaches will briefly be reviewed where different parameters are adapted to speech rate variations.

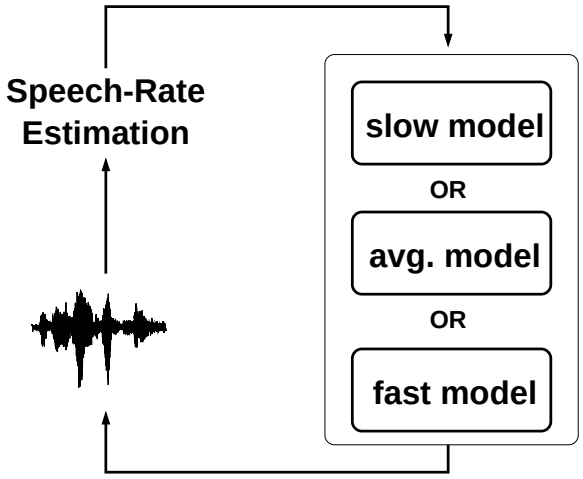

Figure 4.3.: Decoding with rate dependent models. First, a rate estimation is carried out on the test utterance upon which the corresponding model for decoding is chosen. The decoding is then performed with the selected model.

Acoustic models. The adaptation of the entire acoustic models is the most intuitive approach. It affects both, the duration modelling by means of the transition probabilities and the modelling of the spectral characteristics, which are captured by the output probabilities. It can be argued that simple variation in the duration of the segments — provided that the segments are not deleted and still exhibit a sufficiently large duration in order to traverse all states of the model — are captured by the transition probabilities. In contrast, effects on the spectral characteristics of the speech signal, such as reduction, are modelled by the output probabilities.

This approach was followed in experiments reported in [MTÁ98], where three different

sets of rate dependent models were trained. The experiments were carried out on a speech corpus that was created explicitly for the analysis of speech rate variation and consists of three different speech rates for each speaker. These rate-dependent models were used for the decoding and chosen according to an ideal speech rate estimation. In [MTÁ98] relative reductions of up to 64% of the word error rate are reported for slow speech and 19% for fast speech. In total a word error rate reduction of 32% was achieved on the whole corpus. Thus, it is slow speech that profits most from rate adaptation although it exhibits rather homogeneous acoustic characteristic. In contrast, fast speech which is severely degraded profits significantly less from rate modelling. Because of the degradations of fast speech one might have expected another result, namely a larger word error rate reduction for fast speech as compared to slow speech.

However, if the models for decoding are chosen according to a rate classifier, the beneficial influence of rate modelling is completely annulled to a relative reduction of the word error rate of 2.2% for slow speech and increases the word error rate for fast speech by 8.8%.

The high reduction rates that are obtained with an optimal classification of the utterances as fast, average or slow might be due to the fact that the speech corpus was created explicitly for the analysis of speech rate variation. It can be assumed that this corpus exhibits an unnaturally high amount of global speech rate variation with only low local variations. Nevertheless, these results indicate that it is far more difficult to adapt to fast speech than to slow speech.

A more detailed approach was followed in [ZFW+00] and [ZFS00] where the speech rate was computed on words. For each frequently occurring word in the training data two or three versions were trained depending on their relative duration. With this procedure rate dependent phone models were trained. During decoding fast and slow word models were competing against each other so that local speech rate variations could be captured. This approach was able to achieve a significant reduction of the word error rate on a large corpus of spontaneous speech. But compared to the results reported in [MTÁ98] the relative reduction of the word error rate of 1.7% the improvement appears small. However, this might be due to the different corpora. As has already been pointed out, the large improvements reported in [MTÁ98] may be due to the recording situation which might have induced rather artificial speaking styles with the different speech rates where a clear speaking style was chosen and only few local variations of the speech rate occurred.

Additional modelling of pronunciation variants enhanced the relative performance by another 0.3% yielding a total of 2.0% relative word error rate reduction. The modelling of pronunciation variations alone achieved a word error rate reduction of 0.4% which indicates that a modelling on the acoustic and symbolic level is additive but that acoustic modelling has a stronger beneficial effect on the performance than symbolic modelling.

Since fast speech causes a high loss in recognition accuracy, in [MFM95] an adaptation to exclusively fast speech was performed. The adaptation was performed with only 5% of the fastest sentences in order to capture the most characteristic changes in the acoustic properties of the speech signal. On fast speech this adaptation yields a decrease of the word error rate of about 14%. However, on the remaining slow utterances this leads to an increase of the word error rate by about 10%.

To sum up, the beneficial effect on the performance of models adapted to fast speech seems to be limited to a relative word error rate reduction of under 15%.

Transition probabilities. Since the adaptation of the whole model requires a sufficiently large adaptation corpus, many experiments perform the adaptation on a subset of the parameters. Duration as one of the most obviously affected parameters has received most of the attention. For example, manually designed higher exit probabilities for fast speech result in a reduction of 15% of the word error rate [MFM95]. In [VM96] the adaptation of the durational parameters of an MLP yields a word error rate reduction of 5.7% which is slightly better than the adaptation of the transition and observation probabilities in the same experiment. The same effect is reported in [MFM97], where the adaptation of the transition probabilities proved to be better than the modification of the output probabilities. A data driven adaptation of the transition probabilities was reported to reduce the word error rate of slow speech by 13.9% but only by 5.9% for fast speech [MTÁL97].

These experiments suggest, that duration modelling provides a larger reduction of the word error rate than an adaptation of the spectral parameters. However, if the relative error reduction is compared to that achieved by the adaptation of the whole model, this suggestion becomes questionable. While the reduction rates lie around 10% for the adaptation of the durational parameters, the re-training of the whole models achieved error rate reductions of over 20%. From a theoretical point of view the adaptation of the whole acoustic model has a greater beneficial potential than the adaptation of some single parameters only. Since especially the acoustic characteristics of the vowels and consonants are affected by speech rate variations, a modelling of these parameters should provide the greatest beneficial effect. While durational effects can also be severe, they can be captured also by a different strategy where the topology of the whole HMM is restructured. However, there exist only few approaches in this direction.

HMM topology. Closely related to these approaches is the modification of the HMM topology. In order to make use of a continuous rate measure in [TY99] it is suggested to augment HMMs by additional rate dependent output distributions which are attached to the state transitions. Although this would allow the use of a continuous rate measure, in this approach each triphone model consists of two parallel state sequences, one for fast and one

for slow speech. These sequences only differ in their transition probabilities while the states' output probabilities are shared. This topology is chosen in order to prevent a change of the speech rate during the transition of a phoneme. Since no experiments were carried out within this scenario it remains to be shown, whether this approach is superior to the discrete rate modelling approaches.

Scoring Scheme. In [MTÁL97] it is argued that the reliability of the acoustic models decreases with increasing speech rate. From this observation a new approach was derived where the weight of the language model and the word penalty were adapted to the speech rate. This was achieved by setting the weights for three different speech rates. This approach reduced the word error rate by about 15% on slow speech but only by about 1% on fast speech. This approach fits nicely with findings reported in Chapter 3 which showed that the more predictable a word is the less accurate it is articulated and vice versa [Ayl00].

Number of parameters. One of the crucial problems of rate modelling by the training of rate dependent models is the problem of sparse data. In [FPR99] an optimised clustering algorithm is presented in order to address this problem. This algorithm aims at optimising the number of distributions per model by means of a small cross-validation set. The optimal number of prototypes is affected by two restrictions. On the one hand, the more Gaussian densities are estimated, the better the data is modelled. On the other hand, if too many densities are used to model the available data, no generalising power will be left. Thus, by slowly increasing the number of Gaussian densities for fast speech models while evaluating their performance on a cross- evaluation set, an optimal number of Gaussians was achieved. It was observed that vowels receive significantly more Gaussians than consonants with a general decrease of 30% of the entire model size. However, the overall word error rate was only reduced by a relative amount of 3%.

4.2.2. Feature Adaptation

Instead of altering the models, feature adaptation approaches focus on modifying the observed feature vectors in order to better fit the model. The adaptation, which can be performed during both training and recognition, is generally a manually designed rather than a data-driven function.

Like in the model adaptation schemes, first of all a speech rate estimation is performed for each utterance. In a continuous adaptation scheme this value is taken into account as a factor in the following feature modification. Discrete adaptation schemes, i.e. the classification of the rate into discrete rate classes, have not been reported in the literature on feature adapta-

tion. However, only few experiments based on a feature adaptation approach are reported in the literature.

An approach that was able to significantly reduce the error rate on fast speech by modifying the features is reported in [RHAH99]. The fundamental idea of this approach is to adjust the observed length of a phone to its mean duration. This is achieved by a cepstrum length normalisation which is based on a continuous rate measure that estimates the duration of each phone. In order to derive a phone specific normalisation factor, a preliminary recognition pass is required. For each phone the factor for the mean phone duration is computed. In accordance with this factor the phone length is normalised by inserting new feature vectors either by uniformly copying each frame, or by interpolating a new vector by its neighbouring frames. This procedure is performed for each phone with a shorter duration than the average. The intention of this approach is, first, to overcome the problem that short phone realisations are too short to pass through all states of a phone model. Secondly, it is assumed that the dynamics of consecutive vectors are smoothed out by the interpolation of new speech frames. This approach yields an error reduction of about 16.5% for fast speech, when the stretching factor is computed as a mean over the whole utterance. A phone-by-phone stretching yields no achievement because the preliminary segmentation and phone recognition for the computation of the stretching factor are too error-prone. These errors are smoothed out by an averaging of all factors over the whole utterance.

While this approach focuses on the modelling of duration, it also affects the spectral characteristics of speech. By inserting new speech frames the dynamics are smoothed out and the trajectories of the feature vectors over time become less steep.

Since this approach is very sensitive to errors of the preliminary segmentation of the test utterance, which especially occur in spontaneous speech, in [Pfa00] a variant is introduced which makes use of a signal based speech rate estimation where no prior segmentation is needed. In this approach the normalisation factor is based on the mean speech rate measured over a so called *spurt*. Spurts are parts of the utterance where no silence occurs. The measured speech rate of the spurt is then divided by the mean speech rate of the whole training corpus giving an estimate of the relative speech rate. By this approach, errors due to misrecognized phones are avoided. Instead, a smoother scaling of the frame rate is achieved. However, this smoothing increased the performance of the recogniser only little.

4.3. Summary

The survey of the literature on speech rate modelling reveals that most of the approaches follow a similar strategy by separating the training material into discrete rate classes which are then used for the training of rate dependent parameters. During decoding a rate estima-

tion is applied according to which the corresponding parameter sets are chosen. In these approaches the measurement of the rate plays an important role because a wrong estimate causes an inappropriate parameter set to be applied which can lead to higher error rates.

In general the rate is measured as a mean over a whole utterance. It is assumed that the rate does not change too much if the utterance is short enough. However, for corpora consisting of longer utterances shorter stretches of speech are isolated which are bordered by silences. Thus, during decoding only one set of parameters is applied so that no variation of the rate class is allowed in an utterance. Only few approaches allow a change of the models after each word. However, these approaches only yield minor performance increases.

For the classification of an utterance or a shorter stretch of speech as fast or slow generally duration based rate measures which reflect the syllable or phone rate are applied. Only few modelling schemes make use of the acoustic characteristics of the speech signal in order to determine the rate of speech.

The results of all approaches show a similar picture. The highest gain is achieved for slow speech while only little improvements can be observed for fast speech. However, there seem to exist large differences between the corpora as to how much improvements can be achieved. On smaller corpora of read speech generally higher improvements are achieved while the modelling on spontaneous speech corpora only yields minor improvements. This indicates that the approaches are only successful on special corpora.

To sum up, approaches for the modelling of speech rate variations generally consist of an elaborate training or adaptation phase where the models are adapted to a certain speech rate. However, the improvements reported in the literature indicate that the systems are only isolated solutions for a special corpus since the gains for more natural speech remains significantly smaller.

5. Implications for Effective Speech Rate Modelling

The current approaches to speech rate modelling in speech recognition do not represent satisfying solutions. They tend to provide only minor increases of the performance while requiring sophisticated techniques in both training and decoding. A more substantial system is needed, which provides an effective model of speech rate variation that is valid for different speech corpora and imposes only minor efforts on the training and decoding phase. While the creation of such a system lies beyond the scope of this work, fundamental guidelines will be derived by the realisation and evaluation of several rate modelling schemes in this work. This is done by combining acoustic-phonetic methods with techniques from automatic speech recognition. In order to model the underlying effects of speech rate variations explicit knowledge about their nature has to be derived from the speech corpora and combined with data-driven approaches which are appropriate for the probabilistic approaches used in automatic speech recognition. Thus, before a more effective rate modelling scheme can be built a reliable base providing the required information has to be created. The aim of this work is to provide such a fundamental and general base for speech rate modelling.

In the previous chapters the influence of speech rate variation on the acoustic properties of speech and their effects in speech recognition have been discussed. However, while there exist many approaches to model rate variation in speech recognition, no attempt has yet been made to systematically compare the effects occurring in speech recognition with the findings of acoustic-phonetic analyses.

A common approach to deal with rate variations in automatic speech recognition is a classification of the speech data into discrete rate classes. With the data of these classes an adaptation of the model parameters is carried out. However, this means that there exists a parameter set for each rate of speech which is independent of the parameters of the other rate classes. This leads to sparse data problems. Therefore, it would be desirable to make use of the information from the other rate classes or from a systematic relationship between the speech rate and the acoustic features.

Results from acoustic-phonetic analyses show that the acoustic characteristics of speech undergo a systematic transformation from slow to fast speech. Generally, this transformation

is an acoustic reduction of the formant frequencies. This means that the phones become less distinctive in fast speech. But it has also been shown that acoustic reduction is not uniquely related to fast speech. Fast speech does not necessarily induce acoustic reduction. It is possible to avoid reduction by speeding up articulation in order to reach the articulatory and acoustic targets. This speeding up of articulation is reflected in the dynamic characteristics of the speech signal. Reduction can also be caused by other factors such as the lack of linguistic stress or the emotional state of the speaker.

This indicates that there are at least two dimensions at which a change of the speech rate affects the acoustic characteristics of speech. First, there are effects in the temporal dimension of the signal. This implies that less time is available for the pronunciation of a phone or a word. Secondly, the spectral dimension is affected either by reduction or by faster formant movements. Thus, in order to capture the effects of speech rate variation it is necessary to apply different measurements: one for the durational domain and at least one which reflects effects in the spectral characteristics.

In order to investigate these aspects, both analyses of the acoustic properties of a speech corpus as well as recognition experiments are carried out. In the literature different effects of rate variation can be found in different kinds of speaking styles. Therefore, it is necessary to carry out an acoustic-phonetic analysis on the speech corpus which is used for the speech recognition experiments.

5.1. Acoustic Analysis

The purpose of the acoustic analyses is twofold. Firstly, they aim at determining the extent to which reduction or dynamic changes occur and what kind of speaking style is dominating. Secondly, it is investigated which kind of measure describe these effects best. In detail, the following questions are addressed.

1. What speaking style is used by the speakers of the corpus? If a rather informal speaking style is dominating a tendency to reduction especially in fast speech will be observed. On the other hand in a clear speaking style the shorter duration of the segments in fast speech will be compensated by faster movements of the articulators which are reflected in steeper formant tracks. Therefore, measurements of both the static and the dynamic spectral features are taken.

2. Along with this investigation it is analysed which kind of measure is able to capture the expected effects of speech rate variation. This is done by measuring the extent of centralisation and coarticulation.

3. Finally, the proportion of coarticulation as compared to centralisation is measured in order to determine in detail the nature of the shift that the formant frequencies of the vowels undergo from slow to fast speech.

Thus, the acoustic analysis not only aims at revealing the changes of the spectral properties in the signal of fast and slow speech. It also provides measures that can be used for the modelling of speech rate in speech recognition experiments.

5.2. Speech Recognition Experiments

From the results of the speech recognition experiments reported in Chapter 4 and from the insights into the systematic changes of the acoustic features some new questions arise for the modelling of speech rate variation in speech recognition.

1. Does rate modelling based on a reduction measure perform better than a duration based modelling? In general duration or syllable rates are used to capture variations of the speech rate. This measure only focuses on the temporal domain and is not a reliable indicator of the spectral characteristics of the signal. Thus, it is analysed whether the introduction of a further measure which also considers the spectral effects such as reduction and coarticulation is beneficial in the modelling of rate variation.

2. How accurately can the speech rate be modelled? In the presented approaches the rate has generally been classified into two or three classes. Given the large range of rate variation and the different effects it can have, this seems not enough to obtain an optimal model of the speech rate. It is, therefore, an interesting question whether a continuous modelling of the speech rate is possible.

3. Furthermore, the question arises how to determine the speech rate on the test set in order to apply the corresponding models for decoding. In the presented approaches the different parameter sets are generally chosen according to a preliminary rate estimation. However, since there are complex interactions of duration and spectral effects, a more data-driven choice of parameters during decoding is investigated.

4. Is it possible to find an optimal separation of the training data? As the applied measure for the division of the utterances into different rate classes is not an optimal predictor of the acoustic phenomena, a better performance should be achieved by a data-driven approach where the training data is separated according to the classification of the rate-dependent models.

5. How strong is the effect of rate variation as compared to other sources of variation such as speaker variation or environmental noise? One might argue that the differences between speakers are higher than the differences caused by speech rate variation. However, since the effects of rate variation are so highly systematic in their nature it is assumed that the modelling of different speech rates yields a higher generalisation power than the modelling of speakers.

These questions are addressed by systematically investigating the relevant aspects of rate-dependent models. Although rate-dependent models suffer from fundamental shortcomings, such as a classification of rate in discrete rate classes and the lack of the use of the systematics in the change from slow to fast speech they still are a convenient method for detailed analyses of speech rate modelling, since they are able to model both parameters of the temporal and the spectral domain.

6. Acoustic Analysis

Before speech recognition experiments are carried out, analyses of the acoustic properties of vowels in the speech signal are performed in order to better understand what kinds of acoustic variations co-occur with speech rate variation in the corpora at hand.

The analyses are carried out on a large corpus consisting of spontaneous speech gathered from task oriented dialogues which are described in more detail in the next section[1]. Since most of the investigations discussed aboved were carried out on small corpora generally consisting of read speech the current analysis aims at replicating the results on a larger and more natural corpus. In order to do this it was necessary to make use of automatic labelling and segmentation tools. This means also that the measurements of the acoustic features were performed automatically. This approach allows the effects of rate variations to be investigated on a larger data base. Since in the investigations presented above no quantification of the influence of centralisation and coarticulation are made, this study also aims at comparing these two effects.

6.1. Corpus

The current investigation is based on a large corpus of spontaneous speech recorded in the Verbmobil project [KLP+94]. It consists of dialogues with two participants negotiating one or more appointments. Although the scenario is an artificial one set up for the purpose of speech recordings no instructions concerning the speaking style or rate were given to the subjects. The recorded speech is clearly spontaneous (although a button had to be pushed by the speaker to take the floor) and contains the usual features known to occur in spontaneous speech such as hesitations, corrections and a high variability in speech rate. In summary, the corpus consists of around 14,000 utterances from over 650 speakers from different regions of Germany and contains 303,746 words from a lexicon of 6,258 words. The total number of utterances covers a duration of approximately 33 hours of speech. The mean length of an utterance is about 22 words with a mean duration of about 8.6 s. The recordings were performed in an ordinary office environment with close-talking microphones. The speech

[1] Some of the results described in this chapter have been published in [WFS00].

signal was recorded with a sampling rate of 16 kHz.

6.2. Experiments

In order to relate the results obtained from formant analyses to the acoustic features used in automatic speech recognition two sets of experiments were carried out. The first set consists of an analysis of the formant frequencies as obtained from an automatic formant tracker. The results of this investigation are therefore comparable to those reported in the literature. The second set of experiments is based on the mel-filtered spectrum of the speech signal. These rather unusual features for phonetic analysis are chosen because they can easily be computed from the spectrum without the need of an elaborate formant tracking tool which can be error prone. This constitutes a convenient basis for the definition of easy-to-compute measures of acoustic reduction which can be used in automatic speech recognition.

Both analyses are carried out on a segmentation of the speech data achieved by a forced alignment with a standard semi-continuous HMM speech recognition system [Fin99]. The alignment is based on the official transcription as provided by the Verbmobil project. This means that every token of a word is represented by the same phonotypical transcription. Since the alignment is based on a frame-wise segmentation of the signal, segment boundaries can only occur at those points which represent the end of a frame. With a frame rate of 10 ms the segment boundaries in the following experiments occur only at multiples of 10 ms.

While the automatic segmentation allows the analysis of a large corpus of spontaneous speech it exhibits on the other hand certain shortcomings concerning the precision of the labelling and segmentation. For the segmentation a phonotypical transcription of the lexical entries is used. This means that irrespectively of the actual pronunciation each token of a word receives the same phonetic transcription. Therefore, problems can occur when the pronunciation of a word deviates from the phonotypical transcription. For example, the word ”Telefon” (”telephone”) is represented as `[tel@'fo:n]` with the word stress on the last syllable. A regional pronunciation variation is `['tEl@fon]` with the first syllable being stressed. Thus, the segmentation which uses the phonotypical transcription will identify the `[E]` as `[e]`. As is pointed out by Van Bergem in [vB93b], the kind of reduction which occurs in pronunciation variants such as `[tEl@fon]` can be interpreted as lexical reduction, because the speaker intends to produce an `[E]`. In contrast, acoustic reduction is defined as a loss of spectral quality that ”occurs in vowels that were *intended* to be full vowels” ([vB93b, p. 3]). According to this, the produced vowel should be labelled as `[E]` and should not occur in the statistics for the tokens of `[e]`, where it would indicate acoustic reduction.

However, for the current investigation it is not reasonable to follow this distinction mainly for practical reasons. First of all, the investigation is meant to analyse acoustic effects that af-

fect the performance of automatic speech recognition systems. Since the recognition system with which the recognition experiments are carried out does not make use of pronunciation variants it has to deal with both, acoustic and lexical reduction on the level of acoustic modelling. This means that all variants that are produced for the word ”Telefon” receive the same phonotypical transcription `[tel@fo:n]`, so that all tokens of the first vowel are labelled as `[e]`. The model of `[e]` therefore has to capture all of the acoustic variations. Thus, in this procedure the effect that Van Bergem calls lexical reduction cannot be distinguished from acoustic reduction. A more general reason not to follow this distinction between lexical and acoustic reduction is the problematic notion of the word ”intended”. Since it would be a difficult task to determine for each token which phonemes were actually intended by the speaker, a distinction of lexical and acoustic reduction does not seem feasible in the current investigation.

Apart from the labelling and segmentation of the speech data the measurements of the acoustic properties of the speech signal have to be performed. The formant tracking was performed automatically with the ESPS formant tracker[2]. The formant tracker used 20th order LPC-coefficients which were computed every 10 ms on a 16 ms window of speech. The tracker scanned the whole frequency range for five formants from which only two were used for further analysis. Based on the segmentation of the speech signal the following values were computed for all vowels (cf. Fig. 6.1):

1. The duration of the phone as provided by the forced alignment
2. The average F1 and F2 over the whole weighted vowel segment
3. F1 and F2 at the first, middle and last frame of the vowel segment
4. The first derivatives of F1 and F2 at the first, middle and last frame of the segment

The average formant values were computed as a weighted mean over the whole segment in order to emphasise the values in the middle of the segment and attenuate those of the boundaries. With this procedure it was intended to capture the smoothed target values and neglect the contextual influence at the boundaries. The first order derivatives of the formant frequency values are computed as a first order regression over five frames consisting of the intended frame and two frames on each side. They measure the steepness of the formant track over time.

For the computation of the spectrum a similar pre-processing was performed as for the computation of the features for the speech recognition system. This means that a mel-filter

[2] The formant tracker is part of the Entropic Signal Processing System ESPS which is a commercial software package from Entropic Research Laboratory, Inc.

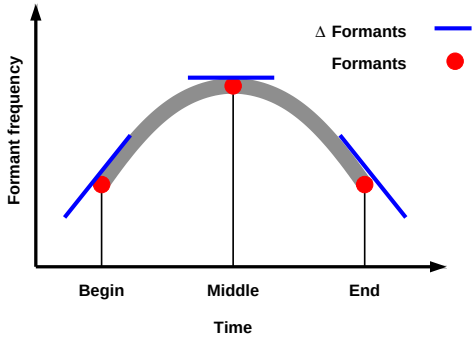

Figure 6.1.: Measurements taken for the vowel analysis at three different points of the segment.

bank was applied on the spectrum which summed the energy of certain frequency bands according to the mel-scale (cf. p. 16). The result of the mel-filter bank is a 32-dimensional vector which represents the energy levels of the mel-spaced centre-frequencies of the resulting channels. Thus, instead of two formants 32 values of the channels of the mel-filter bank as presented in Chapter 2 were computed for each frame. The values for this analysis were computed by exactly the same procedure as for the formants. For the computation of the first order derivatives this means that for each value of the mel-spectrum the temporally surrounding four values of the same spectral band were used.

6.3. Results

Analyses were carried out on the three most distinctive vowels `[a:]`, `[i:]` and `[u:]`. The segmentation produced 38,168 instances of `[a:]`, 23,464 instances of `[i:]` and 7,927 instances of `[u:]`. This inhomogeneous distribution is an effect of the corpus which is caused by the specific task of scheduling an appointment, which elicited a specific subset of the lexicon. For example, the word ”Tag” (”day”) `[ta:k]` is an important word in this domain as well as the discourse particle ”ja” (”well”) `[ja:]` which both contain an `[a:]`. On the other hand there is no such highly frequent word containing an `[u:]`. Thus, the different number of occurrences for the different vowels represents the specific lexical characteristics of the corpus rather than a general tendency in the German language.

For the following analyses the samples of each vowel are divided independently into four duration classes so that each of the classes contains approximately 25% of the data of each vowel. A summary of the mean durations of the four duration classes per vowel is given in Table 6.1.

The mean class durations show that there is a consistent and considerable range with the

		[a:]	[i:]	[u:]
Dur. class	long	225 ms	205 ms	216 ms
	med. long	117 ms	102 ms	114 ms
	med. short	79 ms	69 ms	80 ms
	short	45 ms	41 ms	45 ms
	All	111 ms	98 ms	107 ms

Table 6.1.: Mean durations of vowels in different duration classes.

long mean durations being five times as long as the mean of the shortest class. It should be noted that the lower limit for the segmentation of the vowels is 30 ms because the structure of the HMMs for a long vowel requires at least three frames for the alignment. This is a considerable shortcoming of the automatic segmentation. Also, the mean durations are affected by the intrinsic durations of the different vowels. The open vowel [a:] shows an about 13% longer duration than the closed vowel [i:] for which articulation is faster because the jaw does not have to be opened to such a large degree as for the [a:].

6.3.1. Formants

For a general picture of the behaviour of the formant frequencies of the vowels with different durations the means of the average first and second formants over the whole weighted segment were computed according to their duration class. The mean values are shown in Figure 6.2. As can be seen the vowel space of the vowels with the shortest durations is clearly reduced as compared to that of the longest vowels.

For the statistical analysis of this trend the centre of the vowel triangle of each duration class was computed. As the formant frequencies for each vowel and duration class are approximately normally distributed the distances of the mean values to the centre were computed instead of the distance for each token[3]. The distances of the mean values of the three vowels to the duration specific centres were computed for each duration class. As is shown in Table 6.2 the distance of the mean F1 values to the centre was reduced by 40% and for the mean F2 values by 46%. In analogy, the mean values of the vowel space areas were computed using the means of the vowel formant frequencies for the four duration classes. The results show that the vowel space area of the shortest duration class is reduced by over 70% as compared to the vowel space area of the longest vowels. Thus, both measures for vowel

[3] With this procedure a too strong influence of the variances is avoided which would occur if the distances were computed item-wise. As the variances represent the high variability of the speech sample caused by the large amount of different male and female speakers with individual vowel spaces, the comparison of means was chosen as an abstraction over the speaker specifi cities.

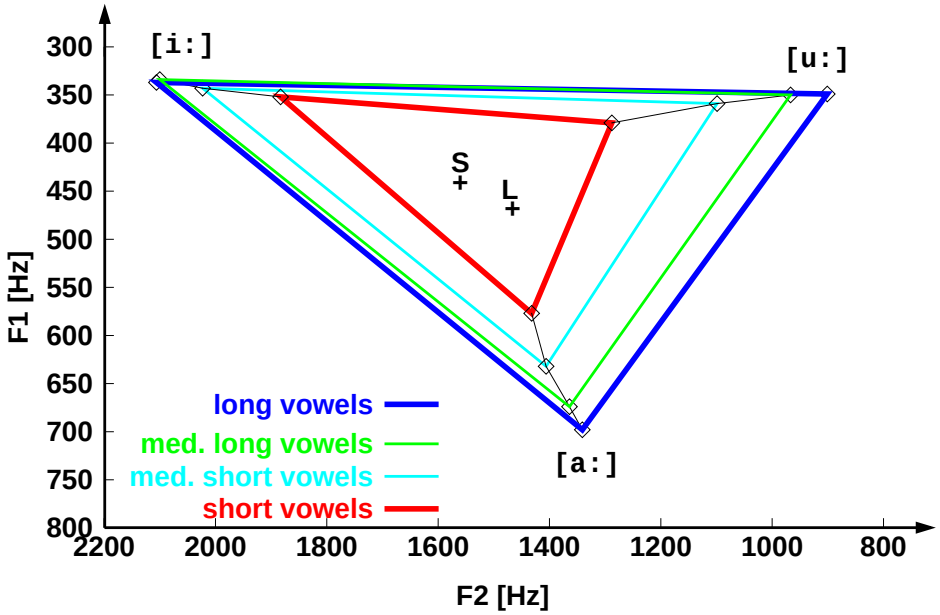

Figure 6.2.: Mean formant frequencies of the vowels [a:] [i:] and [u:] with different durations. Only the vowel space areas of the longest and shortest vowel classes are shown. L denotes the vowel space centre of the long vowels, S the centre of the short vowels.

reduction show a clear reduction of the formant frequencies of vowels with shorter duration.

These results only reflect the effect of duration on the mean formant frequencies over all vowel tokens regardless of their context. However, in order to obtain a clearer picture of the reductional effect of the articulation rate it is important to know to what extent the phonemic context influences the position of a vowel in the vowel space given a certain duration. In other words, does a shorter duration cause a shift of the formant frequencies towards the vowel space centre or rather towards its context which would indicate an increased coarticulation?

Since the Verbmobil corpus consists of spontaneously produced speech no systematic

	Distance to Centre		Vowel Space Area
	F1	F2	
Long vowels	160 Hz	443 Hz	212,912 Hz^2
Short vowels	96 Hz	241 Hz	60,849 Hz^2
Reduction	40%	46%	71.4%

Table 6.2.: Reduction of the vowel space and the distance to the vowel space centre from long to short vowels.

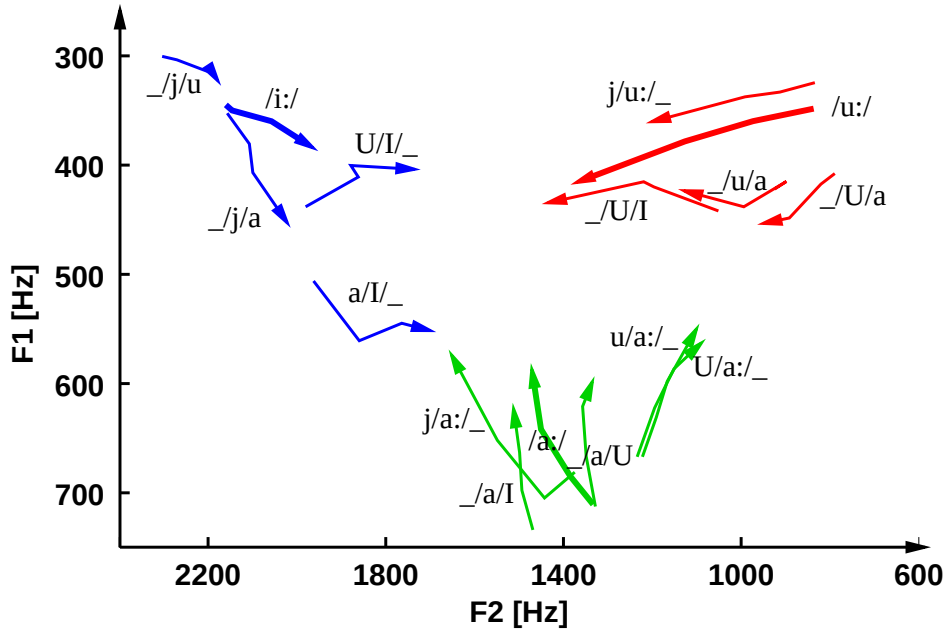

Figure 6.3.: Shift of the formant frequencies of vowels with different phonetic contexts. The arrows show the direction of the shift of the formant frequencies of vowels with shorter durations.

variations of the contexts on the corner vowels can be obtained. Therefore, for this analysis only a few vowel-context combinations are selected which are assumed to exhibit the most severe coarticulatory effects. As the corner vowels `[a:] [i:] [u:]` represent the outer points of the vowel space it is assumed that they are affected most by centralisational and coarticulatory effects. For the same reason they constitute the contexts with the highest bias potential. The context vowels at the outer boundaries of the vowel space are assumed to impose a stronger coarticulatory influence than vowels lying near the centre which would rather impose a centralising influence causing the position of the affected vowels to shift towards the centre. However, in order to measure the extent of centralisation in contrast to coarticulation it is important to analyse those cases where the effects can be expected to be most divergent. Therefore, the vowels `[a:] [i:] [u:]` and some of their lax or neighbouring counterparts are selected as context phonemes. Altogether, 14 combinations of two vocalic segments which meet these specifications were analysed (cf. Table 6.3).

From these vowel and context combinations the mean formant frequencies were computed for the four different duration classes which contained 25% of the biphone tokens. The results as shown in Figure 6.3 indicate that there indeed exists a strong increase in the coarticulatory effect with decreasing vowel duration. For example, the `[a:]` in the `j/a:/` biphone shows a clear tendency towards `[i:]`. The very short `j/a:/` tokens have almost the same

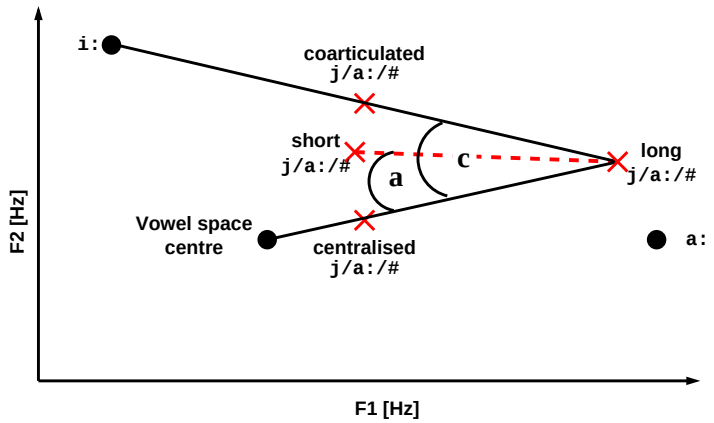

Figure 6.4.: Measurements for the quantification of the centralisation versus coarticulation effects.

formant values as the [I] in the diphthong a/I/. However, there exist also some centralisation effects. The [a] in /a/I, for example, is almost shifted on a straight line towards the centre although it is a diphthong which is assumed to contain rather more coarticulation than monophthongs.

In order to quantify these effects the centre of the vowel space was computed as the mean across the mean values of [a:] [i:] [u:]. Based on these mean values the angle between the direct line from the means of the vowel tokens with the longest durations towards the centre of the vowel triangle which represents the net centralisational effect and the direct line towards the ”target” frequencies of the idealised context was computed (cf. Figure 6.4). This line represents the direction of the net coarticulation without any centralisational tendencies. This angle c describes the whole range between net centralisation and net coarticulation.

The direction of the shift which the formant frequencies undergo with decreasing duration is plotted with the red dashed line. The angle a describes the angle between this line and the centralisation direction. Thus, the ratio $\frac{a}{c}$ describes the relative amount of coarticulation. If $a = c$ the shift consists of 100% coarticulation. If $a = 0$ there is no coarticulation. These ratios were computed for all 14 biphones (cf. Table 6.3).

In order to obtain a single value for the whole data the mean over all ratios was computed by weighting all values with the number of occurrences of each biphone. This overall mean ratio shows that the analysed biphones undergo a shift which consists by about 60% of coarticulation and 40% of centralisation as defined by the ratio of the angles a and c. This indicates that reduction as measured by the mean distance of the observed vowels to the centre consists to almost equal parts of both coarticulatory and centralising tendencies. This implies that both effects should be taken into account for the modelling of reduction by

Biphone	a	c	$\frac{a}{c}$	#
j/a:/_	34.08	33.04	1.03	1682
_/a /I	9.87	50.18	0.19	6138
_/a /U	-37.28	94.63	0.39	1934
u/a:/_	102.48	107.57	0.95	146
U/a:/_	109.98	108.81	1.01	145
j/u:/_	5.83	10.90	0.53	372
_/U /I	3.53	2.97	1.19	40
a/U /_	19.89	29.37	0.68	2136
_/u /a	4.00	-27.08	-0.15	130
_/U /a	-10.62	-22.05	0.48	125
_/j /u	0.31	9.88	0.03	372
_/j /a	-26.29	-19.97	1.32	1684
a/I /_	15.12	27.58	0.55	6136
U/I /_	10.66	0.29	36.8	41
Mean			0.61	

Table 6.3.: Angles of the shift of biphones representing the amount of coarticulation caused by decreasing duration (cf. Figure 6.3).

modelling a shift towards the centre as well as a shift towards the phonetic context.

Apart from the static characteristics the behaviour of the dynamics was investigated in some detail. For the analysis of the changes in the formant movements the means of the formant frequencies measured at the beginning, middle and end of the segments were computed. When the formant frequencies of the vowels of the different duration classes are plotted against their duration as shown in Table 6.1 and connected with straight lines the movements as shown in Figure 6.5 can be deduced. As can be seen while the on- and offset frequencies remain almost invariant the target values in the middle of the segments are clearly reduced for the shorter segments. This indicates that the total amount of reduction as seen in Figure 6.2 is mainly due to the middle segment.

The connections of the measured formant frequencies suggest that the rate of change remains stable over the different duration classes. Only for the very short vowels the movements between the first and last frame tend to approximate a straight line which represents a deviation from the direction of the movements in the longer vowels. The first derivatives computed at the same points of the vowel segments show that the steepness of the formant movements at least for the middle and last frames remain relatively stable over the different duration classes (cf. Table 6.4). The values of the derivatives of the long vowels indicate that at the beginning of the segments the movements tend towards a certain target which seems

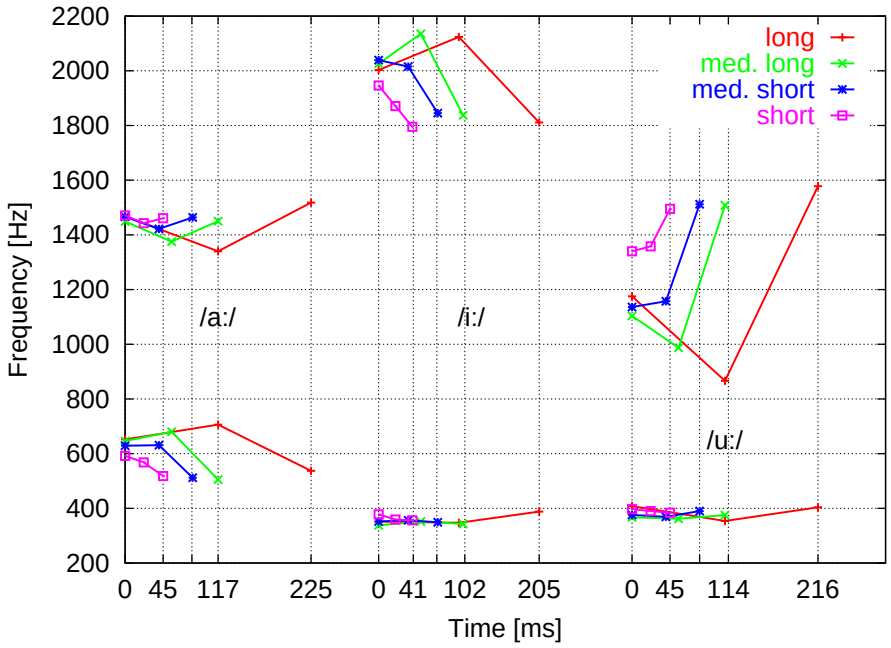

Figure 6.5.: Formant movements of the vowels [a:] [i:] and [u:] as deduced by the measurements of the first and second formants at the begin, middle and end of the vowels. The duration is chosen according to the mean duration of the corresponding duration class. The lower tracks represent F1, the upper tracks F2.

to be reached at a quasi stationary state in the middle of the segment where almost no movement can be observed. In the final part the movements show the opposite direction towards a neutral position. In contrast, the derivatives for the short vowels do not show such a clear separation of the phases. The values for the middle frames indicate that no stationary state is reached but that the movements are continued over the whole segment. However, the values of the derivatives of the second formants of the initial frames of all three vowels indicate a flatter movement for the shorter vowels.

6.3.2. Spectrum

These results only take into account the frequencies with the highest energies, the formants. However, this completely ignores the rest of the spectrum. Although it is generally assumed that most information about the vowel identity is conveyed by the formant frequencies it is known that the spectrum also conveys information such as speech effort or specific speaker characteristics. These informations can be helpful in determining the speech rate. In order to analyse the mel-spectrum a similar approach as for the analysis of the formants was taken.

		ΔF1		ΔF2	
		long	short	long	short
[a:]	Begin	28	8	-41	-18
	Middle	-4	-16	5	-5
	End	-16	-15	25	21
[i:]	Begin	-6	-17	63	-10
	Middle	3	-8	-17	-43
	End	7	16	-24	-35
[u:]	Begin	-32	-40	-128	-38
	Middle	3	-12	23	27
	End	18	17	83	78

Table 6.4.: Mean values of the first derivatives of F1 and F2 at three different points of the vowels [a:] [i:] and [u:]. Only the values of the longest and shortest duration classes are shown.

However, since the spectrum exhibits a highly speaker specific structure an averaging over all speakers would destroy the vowel specific characteristics. For example, the formant frequencies of the same vowel can vary widely between different speakers. This does not severely affect the analysis of formants because the averaging still returns satisfying results. However, in the average spectrum of a vowel over different speakers the formant structure would be lost because the formant peaks at the different frequencies for the different speakers would simply flatten the spectral structure. Therefore, the means of the spectrum were computed speaker-specifically.

Figure 6.6 shows the mean mel-spectrum of the vowel [i:] produced by speaker HAH from the Verbmobil corpus. The values over which the means are computed were measured at the middle frame of all vowel tokens. The spectra depicted in Figure 6.6 show the means of the vowel tokens with the longest and shortest durations.

From the results of the formant analysis which shows a strong reduction of the formant frequencies one might expect a strong effect of the reduced formant frequencies on the whole spectral structure. However, this is not the case. The differences in the positions of the spectral peaks are almost negligible. While the graphs show that the reduction of the formant frequencies does not have a strong effect on the overall spectral structure there are several large differences in the spectra of short and long vowels. First of all, the spectral tilt of the mel-spectra of the short vowels is more level than that of the long vowels. In Chapter 3 this effect has already been introduced. There, the centre of gravity of the spectrum as a measure of the slope of the power spectrum was lower for shorter vowels indicating a steeper tilt of the spectrum. Due to the mel filtering this effect is reversed. However, the phenomenon

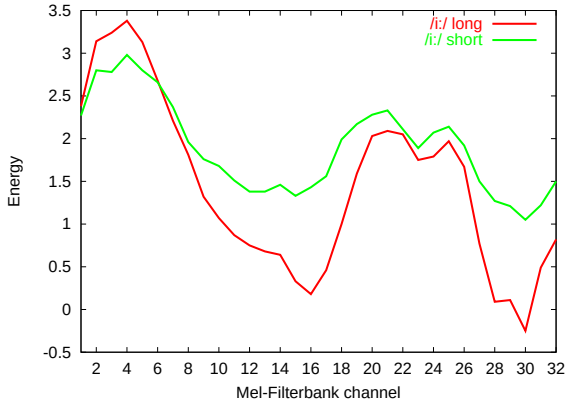

Figure 6.6.: Mean mel-spectrum of the speaker HAH for the vowel [i:]. The means of the tokens with the longest and shortest durations are shown.

remains the same, namely the shift of the global spectral energy in vowels with a shorter duration. Secondly, the formants themselves appear less pronounced in short vowels. In order to capture both effects in one measure the spectral variance was computed over the spectra of this speaker according to the following formula:

$$VAR = \frac{1}{n}\sum_{i=1}^{n}(\mu - E_i)^2 \tag{6.1}$$

where n is the number of frequency bands of the mel-spectrum, μ is the mean spectral energy and E_i is the energy of the i-th frequency band.

The spectral variance was computed for the vowels [a:] [i:] and [u:] for the speaker HAH. The results as shown in Table 6.5 indicate that this measure is highly vowel specific. While for [a:] there is a significant difference in the spectral variance between long and short vowels this effect does not occur for the vowels [i:] and [u:]. This result indicates that this measure is not suitable for all vowels. However, since the effect for [a:] is so clear the measure might be suitable for a classification of a stretch of speech where only a smoothed value has to be taken into account.

Apart from differences in the static characteristics of the mel-spectrum interesting effects can be observed in the behaviour of the spectrum over time for vowels with short and long duration. A characteristic picture of the dynamics of the different frequency bands is drawn by the first order derivatives of the mel spectrum as depicted in Figure 6.7.

Again, the red line shows the values obtained at the first frame, the green line the values at the middle frame, and the blue line those at the last frame. Thus, the red line for the

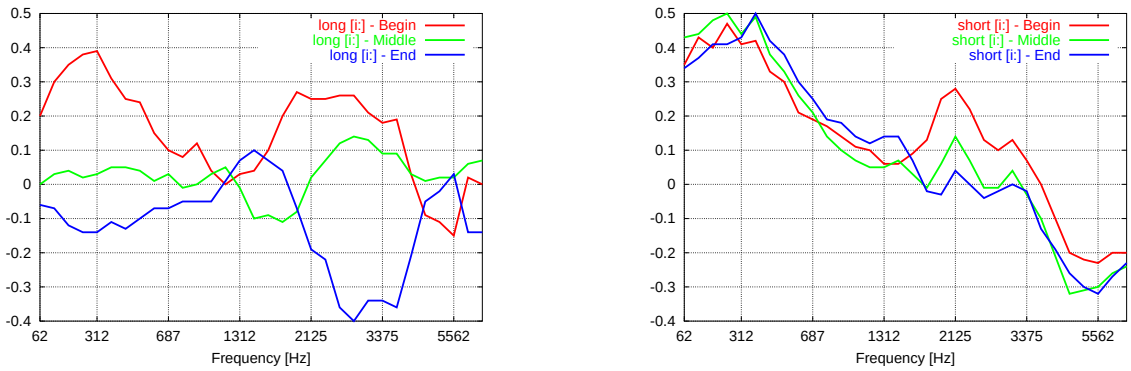

Figure 6.7.: First order derivatives of the mean mel-spectrum of the speaker HAH for the vowel [i:] as measured at the first, middle and last frame of the vowel tokens.

long vowels indicates that at the beginning of a vowel especially the energies of the formant frequencies are rising. In contrast, in the middle of the vowel almost no movements can be observed which indicate a quasi stationary part. At the end of the vowel an inverse picture of the beginning is obtained with the formant frequency energies being decreased. In summary, the derivatives of the mel-spectrum of the long vowels indicate three different states in the articulation. The first phase consists of an on-glide where the articulators are moved towards the target position. A quasi steady state is reached in the middle of the segment and the segment ends in an off-glide phase where the articulators are moved towards the position of the following phone.

These three phases are completely absent in the derivatives of the short vowels. There, the movements remain the same during the whole duration of the vowels.

Vowel	Short Vowels		Long Vowels	
	Mean VAR	Variance VAR	Mean VAR	Variance VAR
[a:]	1.06	0.29	1.48	0.25
[i:]	1.24	0.42	1.26	0.57
[u:]	1.77	0.32	1.63	0.97

Table 6.5.: Means and variances of the spectral variance VAR for the vowels [a:] [i:] and [u:] for the speaker HAH.

6.4. Summary

The results of the acoustic analysis of the German Verbmobil corpus of spontaneous speech in general support the results obtained on smaller corpora as reported in the literature. The reduced formant frequencies in the shorter vowel segments indicate a rather informal speaking style where no special effort is taken by the speakers to produce clearly articulated utterances. Accordingly, the formant movements over time remain relatively stable over the different speech rates. The only difference here is the lack of a quasi steady state in the short vowels as compared to vowels with longer durations.

These results are mirrored in the analysis of the dynamics of the mel-spectrum where a clearly slowed movement was found in the middle of the long vowels which was absent in the short ones. The analysis of the static spectral values reveals that there are significant differences in the spectra of short and long vowels which can be captured with very simple measures of the centre of gravity and the variance of the spectrum.

An investigation of the effects of centralisation and coarticulation indicates that the formant frequencies of the vowels are not only shifted into the direction of the formants of their neighbouring vowels as would be the case with increasing coarticulation. It has also been shown that there is a strong tendency of the formants to shift towards the centre of the vowel space. Thus, the transformation that the formant frequencies of the vowels undergo from slow to fast speech consists in fact of two underlying shifts.

7. Rate Dependent Models

As has been shown in the previous chapters there exists a discrepancy between the phonetic knowledge of rate variation and rate modelling in automatic speech recognition. While it is evident that changes in speech rate cause systematic changes in the acoustic features these systematics are not exploited in approaches with rate dependent models. In order to investigate in which way these systematics could be incorporated in speech rate modelling a series of experiments was carried out on rate dependent models.

In these experiments the following questions were raised.

Firstly, the question was addressed whether a separation of the training data from the beginning is superior to a mere adaptation of general models. In other words, it was analysed whether rate dependent models need a more general basis or not. If the effects of rate variation are severe and cause strong deviations from normal rate characteristics a separation of the models from the beginning should perform better. However, if these effects cannot be captured reliably because of sparse data problems a more general basis of the models is needed.

Secondly, the kind of measure upon which the division of the training data is carried out is analysed. The measure is an important variable in the modelling of speech rate since it determines on what effect adaptation is actually performed. As has been shown in chapter 3 speech rate variation is generally accompanied by acoustic reduction. However, reduction does not always occur with faster speech rate. One can argue that speech rate is not the optimal criterion upon which adaptation should be performed. A criterion that measures directly the acoustic effect of reduction may perform better.

Thirdly, the number of rate- or reduction-dependent model sets is a critical point. While a higher number of rate classes on the one hand dramatically reduces the amount of training data that is available for each class of rate dependent models, this can on the other hand increase the modelling accuracy. Although most rate modelling approaches use around three rate classes, a more detailed modelling strategy has not yet been analysed.

The selection scheme for the models during the recognition phase is another critical parameter. Generally, much effort is taken for the online estimation of the speech rate during decoding. However, the chosen measure may not be the optimal predictor for the kind and

amount of acoustic variations. Also, the rate estimation is an error prone task which introduces further uncertainty into the choice of the appropriate model. Therefore, a model driven selection scheme was analysed in detail.

For the same reason that a chosen measure may not be the optimal predictor of the acoustic effects a different scheme for the separation of the training data was applied. In order to find an optimal separation of the training corpus, a data-driven selection which is based on the application of the rate-dependent models was carried out.

Finally, a more basic question was addressed by comparing rate adaptation with speaker adaptation. Since rate is assumed to be a highly speaker specific variable it is an important question if the applied rate adaptation schemes perform implicitly a speaker adaptation or if rate modelling shows a better generalising power[1].

7.1. Rate and Reduction Measures

In the following investigations a total of eight measures was analysed from which three are durational measures, three are based on formant values and two are based directly on the mel-spectrum. For the analysis the measures were computed on both the training and the test set. For those measures which require a segmentation of the data this can be interpreted as a cheating mode because the segmentation of the test data was achieved by a forced alignment of the phonotypic transcription. It is an interesting question whether this cheating is helpful for the decoding phase.

The first three measures simply give the mean duration of certain segments over a stretch of speech. In the first corpus which consists of read speech with very short utterances this means the whole utterance. In the spontaneous speech corpus where longer utterances are dominating shorter runs are defined according to [GD75] as stretches of speech that occur between two pauses. The duration based measures are defined in a straightforward fashion. VDUR gives the mean vowel duration, DUR the mean segment duration and SYL the mean syllable duration of an utterance. For all measures pauses are excluded. The syllable and segment duration are inversely related to the syllable and phone rate respectively. The vowel duration was measured additionally because vowels are affected most by rate changes. It was expected that a focus on the most rate sensitive segments would capture rate changes even better than other measures.

In order to determine the formant based measures the ESPS formant tracker computed the frequencies of the formants on each frame. For the measurement of the reduction BRED and coarticulation BCOA the formant frequencies f in Hz were transformed to Bark values

[1] Some of the following results are published in [WFS01a] and [WFS01b].

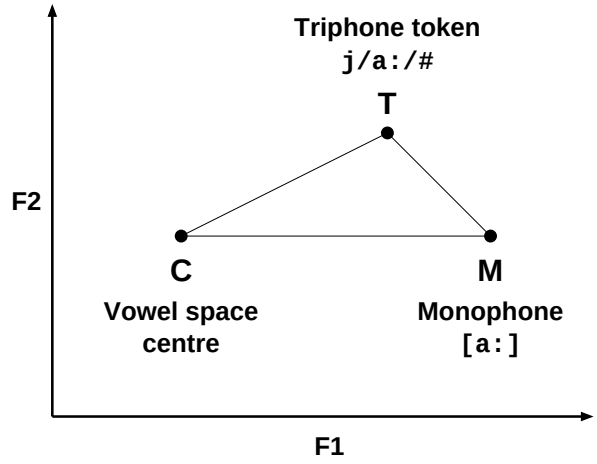

Figure 7.1.: Monophone, triphone and the centre of the vowel space as base for the computation of reduction and coarticulation.

z according to the following function [ZF99, p. 164]:

$$z = 13 \cdot \arctan(\frac{f}{1\,000 \text{ Hz}} \cdot 0.76) + 3.5 \cdot \arctan(\frac{f}{7\,500 \text{ Hz}}) \tag{7.1}$$

In accordance with [Ayl00] the reduction BRED is measured as the Euclidean distance of a token T to the centre C of the vowel triangle normalised by the distance of the corresponding monophone M, which is the mean over all vowel tokens T_i of the vowel (cf. Figure 7.1):

$$BRED = \frac{\|C - T\|}{\|C - M\|} \tag{7.2}$$

Accordingly, the amount of coarticulation is measured as the Euclidean distance of a token T to its corresponding monophone M. For practical reasons this distance is also normalised by the distance of the monophone to the vowel centre. This is based on the fact that the nearer a vowel class lies to the centre, the more likely it is to overlap with another vowel class. Thus, the nearer a vowel class is to the centre the more severe is the effect that coarticulation has on the separability of the vowel classes.

$$BCOA = \frac{\|M - T\|}{\|C - M\|} \tag{7.3}$$

In order to measure the coarticulatory strength of vowels on consonants the slope difference of the F2 trajectories was computed as suggested by Van Son et al. [vSP95] [vSP99]

[vSP96]. For this, the first order derivatives of the second formant $\Delta F2$ were computed over a window of five frames corresponding to 50 ms. These values were determined for the second formant at the end $\Delta F2_{End}$ and beginning $\Delta F2_{Beg}$ of each consonant. The difference gives the slope difference F2SD:

$$F2SD = \Delta F2_{End} - \Delta F2_{Beg} \tag{7.4}$$

These formant based measures are rather elaborate because they are difficult to compute online since the formant tracking generally is a time consuming and error prone process. From this point of view computationally less elaborate measures are more interesting. This is the case for the centre of gravity of the spectrum COG which can be computed directly from the spectrum and is a measure of the speech effort. In accordance with the acoustic analyses in Chapter 6 the COG is computed on the energy values of the frequency bands as derived from the mel-spectrum:

$$COG = \frac{\sum_i f_i \cdot E_i}{\sum_i E_i} \tag{7.5}$$

where f_i is the centre frequency in Hz of the i-th frequency channel of the mel filter bank and E_i the according energy. Based on the results of the analysis of the spectral characteristics of fast speech in the Verbmobil corpus the spectral variance has been included into the pool of reduction measures:

$$VAR = \frac{1}{N} \sum_{i=0}^{N} (\mu - E_i)^2 \tag{7.6}$$

where E_i denotes the energy of the i-th energy band and N the number of all energy bands of the whole mel-spectrum. μ is the mean energy over all frequencies of a frame.

These eight measures were computed on all utterances of the SLACC corpus and used independently of each other for the following experiments. For the experiments on the larger Verbmobil corpus only the VDUR was computed.

7.2. Experiments on the SLACC Corpus

The following experiments are based on a different corpus than that of the acoustic analysis. Since the experiments on rate dependent models require many iterations of training and

evaluation phases a smaller corpus was selected upon which many sets of experiments could be carried out. However, this smaller corpus differs in some aspects from the Verbmobil corpus, for example in the speaking style. While the Verbmobil corpus consists of spontaneous speech the utterances from the SLACC corpus are read. In order to analyse the differences between the effects of rate variation in these different corpora some experiments will also be carried out on the Verbmobil corpus.

7.2.1. Corpus

The experiments were carried out on the SLACC corpus (Spoken LAnguage Car Control) which consists of read utterances recorded in a car environment [Sch01]. The utterances contain instructions that are likely to be used for the control of non safety-relevant functions in a car such as mobile phone or air-conditioning. The utterances start with the number of the turn and consist of infinitive constructions such as ”Radio einschalten” (”Turn on the radio”). Thus, with 5.7 words per utterance for the training set and 5.3 words for the test set the mean utterance length is very short.

	Speakers	Utterances	$\frac{Words}{Utterance}$
Training	18	9,207	5.7
Test	4	1,787	5.3

Table 7.1.: The test- and training set of the SLACC corpus.

For the recordings the speaker was sitting on the passenger seat with a microphone mounted to the front jamb at the front shield of the car. The recordings were performed while the car was driving. In total 22 speakers were recorded from which 18 were selected for the training set and four for the test set.

7.2.2. Baseline System

The baseline system is a semi-continuous Hidden-Markov-Model recogniser realised within the ESMERALDA speech recognition environment [Fin99]. The codebook consists of 512 Gaussian distributions described by the diagonals of the covariances only. The word modelling is based on triphones which consist each of a left-right model with two or three states where no skips are allowed. The lexicon of 650 words requires 1400 triphones. After a clustering procedure 1600 different states are established. The training consists of several iterations of the Baum-Welch-algorithm. During decoding a Viterbi beam search is performed on

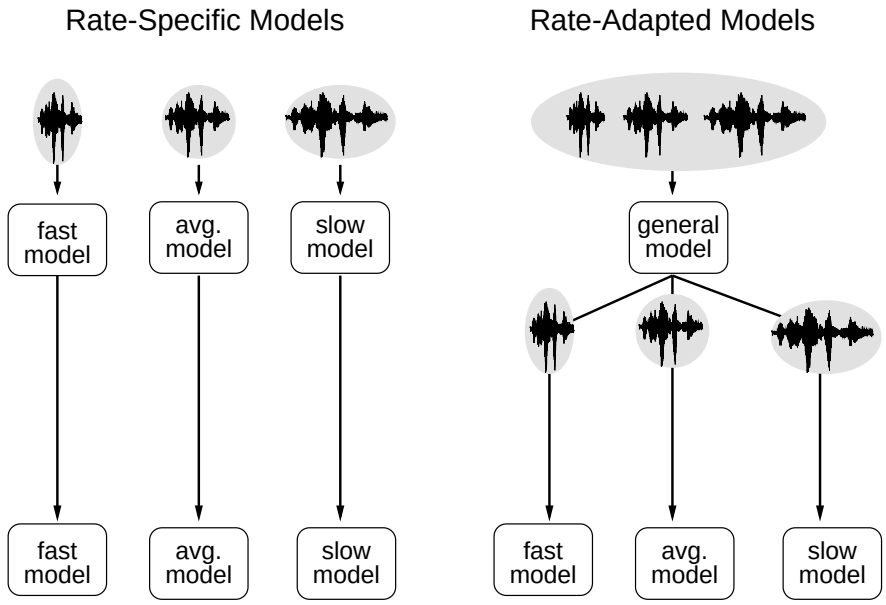

Figure 7.2.: Separation of the training data for rate specific versus rate adapted models.

the lexicon. Since no language model is applied a word penalty is used in order to avoid a dominance of short words.

7.2.3. Basis of the Models

In order to compare the performance of highly specialised models to models which still contain some information of more general speech data two different kinds of models were trained. For the training of the rate specific models only rate specific data was used by separating the training data right from the initialisation (cf. Figure 7.2). Thus, all rate specific model classes contained a distinct set of information. In contrast, the rate adapted models are first trained with the whole data until an optimal performance is reached. Only then a rate-adaptive re-estimation of the model parameters is performed with the according subsets of the training data using the Baum-Welch algorithm.

During decoding all three models were applied and produced scored hypotheses for each utterance. Thus, for each utterance three hypotheses were produced. The hypothesis with the best score was chosen as the recognition result.

In order to analyse the effects of rate modelling on the different speech rates in the test data the test utterances were divided into fast, average and slow utterances for evaluation[2].

[2] In order to obtain comparable results the subsets referred to as slow, average and fast are always the same. They are obtained by computing the mean vowel duration of each test utterance as given by the

	Testset			
	slow	avg.	fast	all
Baseline	20.1	21.5	31.6	25.0
Rate-adapted	17.9	20.2	30.4	23.4
Rate-specific	19.5	22.0	32.6	25.3

Table 7.2.: Word error rates of rate-specific models vs rate-adapted models (VDUR, 3 classes, best score).

The results of the baseline system show that the word error rate is highest for the fastest utterances with 31.6% and lowest with 20.1% for the slowest utterances. In total the baseline system yields a word error rate of 25%.

Table 7.2 shows that the rate-adapted models improve the word error rate of the baseline system up to 23.4% which represents a reduction of the word error rate of 6.4%. This improvement is achieved in all rate classes similarly after only one adaptation step. Further iterations of this adaptation yield roughly the same results. In contrast, the performance of the rate specific models with a word error rate of 25.3% was slightly but not significantly worse than that of the baseline system.

A closer analysis of the test data reveals that the rate specific models perform particularly badly on the fast utterances while for slow utterances the performance can be slightly enhanced.

Thus, especially the fast models benefit from a more general base for the rate dependent models. This indicates that the models of the fast utterances are more heterogeneous which corresponds with the results of the formant analysis of the Verbmobil corpus where the formant frequencies of the fast vowels were not only seen to be reduced but also to exhibit a higher variance.

7.2.4. Rate- and Reduction Measures

In order to investigate the effect of different measures on the performance of rate-dependent models, the criterion according to which the training data is divided into specialised subsets was changed. Thus, for each measure a different division of the training data was obtained.

segmentation of a forced alignment. The margins of the classes are defined by the ranges of the classes established for the training data. The fast test subset contains 829 utterances, the average 451 and the slow subset 507 turns. The range of the mean vowel duration for the fast subset of the training data lies between 32 ms and 65 ms, for the average subset between 65 ms and 78 ms and for the slow set between 78 ms and 187 ms.

	Reference		Duration based rate measures			Formant based rate measures			Spectrum based rate measures	
	Base	RAND	DUR	VDUR	SYL	BRED	BCOA	F2SD	COG	VAR
all	25.0	24.9	23.6	23.4	23.4	23.5	23.2	23.9	24.3	24.7
fast	31.6	32.2	30.5	30.4	29.8	30.2	30.2	30.9	31.1	30.9
avg.	21.5	21.6	20.3	20.2	20.6	19.5	18.9	19.9	20.4	21.5
slow	20.1	19.2	18.3	17.9	18.3	19.0	18.5	19.2	19.7	20.2

Table 7.3.: Word error rates of rate-adapted models with different measures (3 classes).

Therefore, eight different kinds of measures were computed from which each was the base for a set of experiments. In order to divide the training set the mean value of the analysed measure was computed over each utterance. According to this measure the training set was divided into subsets containing the same number of utterances.

Again, all three rate dependent model sets were applied for decoding and only the best scoring hypothesis was chosen for evaluation. Table 7.3 shows the results of the rate adapted models based on different measures with three different classes each. As a comparison a random pseudo-measure RAND has been introduced to test whether potential improvements are due to some artifacts related to the training and test procedure. For example, the larger number of parameters could be a simple explanation of an improved modelling irrespectively of the underlying separation of the training data and the phenomena that are intended to be captured by this approach. However, with a word error rate of 24.9% the RAND condition produced the same result as the baseline system. In general, the formant and duration based measures all provide significant improvements of the performance with a word error rate of under 24%. In contrast, the performance of the models based on the spectrum based measures COG and VAR is only slightly but not significantly better than that of the baseline system.

It should be noted that the best performance for the duration based measures DUR, VDUR and SYL is achieved after one or two training iterations. This is also the case for the not very successful spectrum based measures COG and VAR. However, the performance maximum of the formant based measures BRED, BCOA and F2SD is reached after around ten Baum-Welch iterations. This indicates that the adaptation of reduction involves more complex transformations than the adaptation of duration.

Given the similar results of the duration based and formant based measures the question arises how different these measures are and if they evoke any differences in the division of the training set. The permutation matrix of the duration based measure VDUR and the formant based measure BCOA (cf. Table 7.4) shows how many utterances that are classified as fast with VDUR are classified as fast, average or slow with BCOA and so on. The numbers show that although the utterances are not randomly distributed more than half of the utterances

		BCOA		
		fast	avg.	slow
VDUR	fast	1405	1087	515
	avg.	885	1060	1062
	slow	718	860	1430

Table 7.4.: Confusion matrix of the training utterances classified as fast, average and slow based on the VDUR and BCOA measures.

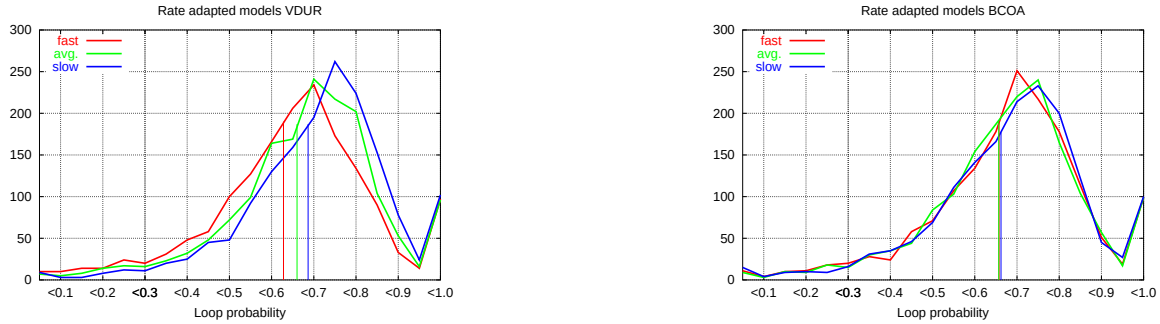

Figure 7.3.: Histograms of the loop probabilities of the rate adapted models with the best performance based on the rate measures VDUR and BCOA.

classified as fast with the VDUR measure are not classified as fast with the BCOA measure. For example, there exist 515 utterances which exhibit a short vowel duration but only few coarticulatory effects as captured by the BCOA measure. This shows that a fast speech rate as reflected in shorter vowel durations is not always accompanied by stronger coarticulation or reduction as discussed in chapter 3. A similar picture is drawn by the remaining duration classes. This shows that there is a significant difference in the composition of the rate specific training subsets.

Based on these figures one might suspect that the coarticulation adapted models BCOA actually perform a duration adaptation, only more slowly. As has already been mentioned the models adapted by the formant based measures need more iterations of the parameter re-estimation. This could indicate that BCOA is simply a worse measure of duration than VDUR itself. In order to investigate this it is helpful to look directly at the adapted parameters of the different model sets. Since it would be a difficult task to analyse the 512 mixture weights of the semi-continuous HMMs together with the corresponding classes of the codebook, the transition probabilities were analysed for this question. This is an easier task since the states of the applied HMMs only have two transition probabilities, namely the loop and the exit probabilities. Figure 7.3 shows the histograms of the loop probabilities of the models based on VDUR and BCOA.

As can be clearly seen the distributions of the loop probabilities for the slow VDUR models are shifted towards higher values than the loop probabilities for the fast models[3]. This is to be expected since a higher loop probability models a longer segment duration. In contrast, the adapted models based on the BCOA measure do not show any significant difference in the loop probabilities of the fast and slow models. Thus, while the rate dependent models of the VDUR measure are clearly adapted to the different durations the BCOA models are not. This means that the BCOA models are obviously adapted to something different than duration. It is therefore an interesting question if the effects of the different adaptation schemes are additive. In order to investigate this question the VDUR and BCOA models were applied together during recognition which means that six models were applied producing six hypotheses for each utterance from which the best scoring one was chosen. However, this combination only achieved a slight decrease of the word error rate to 23.1% as compared to 23.2% and 23.4% for the BCOA and VDUR condition respectively. Thus, although the duration based measures capture different aspects of acoustical degradation than the formant based measures a combination of both measures is not beneficial at the level of hypotheses.

In order to better understand the effects of the rate dependent models on the different rate classes a closer look on the performance on the fast, average and slow subsets of the test utterances is helpful (cf. Table 7.3). Most of the word error rate reduction is achieved on the average rate subsets of the test utterances. The highest reduction is achieved by the coarticulation adapted models BCOA for the average rate subset from 21.5% to 18.9%. This corresponds to a word error rate reduction of 12%. For slow speech the highest error reduction lies around 11% which is achieved by the VDUR models in the best score condition. These results reflect the general tendency of the formant based measures BRED, BCOA and F2SD to better adapt an average rate while the duration based measures seem to perform a better adaptation on slow speech. However, for fast speech the reduction of the word error rate does not exceed 5.7% in the best score condition. Here, the best results are achieved by models adapted to the syllable duration SYL.

In order to get a better understanding of the effects which are captured by the different measures the formant frequencies of the training data for the VDUR and BCOA subsets were computed. The results as depicted in Fig. 7.4 and Fig. 7.5 are based on the formant frequencies as measured in the frame at the middle of the vowel segment. The figures show the mean values for the vowels `[a:] [i:] [u:]` taken from utterances classified as fast or slow according to their mean vowel duration. The ellipses show the standard deviations for each distribution.

[3] There is a high number of loop probabilities near the value 1.0. This is an artifact caused by the chunks on which the training is performed. At the end of these chunks no exit transitions are allowed which leads to loop transition probabilities near 1. Since such chunks generally consist of whole words there exist quite a lot of states occurring at the ends of a word which are affected. However, these values are not affected by rate modelling and can be ignored.

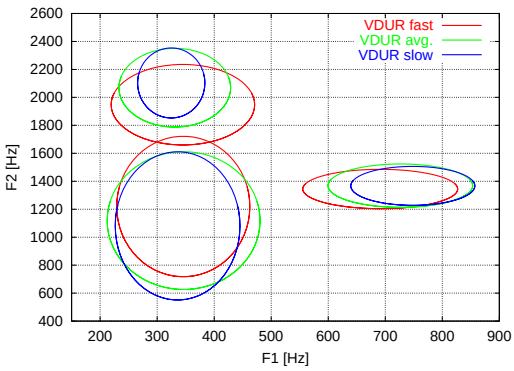

Figure 7.4.: Mean formant frequencies and standard deviation of the vowels [a:] [i:] [u:] in the training subsets defined by VDUR.

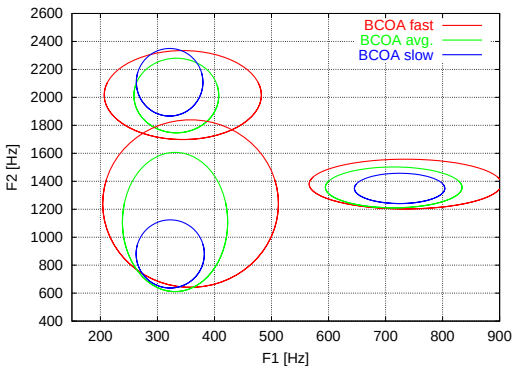

Figure 7.5.: Mean formant frequencies and standard deviation of the vowels [a:] [i:] [u:] in the training subsets defined by BCOA.

While the formant frequencies of the vowel tokens that are used to train the slow BCOA model only show little variances – at least for [a:] and [i:][4] – the range of the strongly coarticulated vowel tokens is greatly expanded. This is exactly the coarticulatory effect that it was intended to model with the BCOA measure. In contrast, the variances of the vowel subsets selected for the training of the VDUR models are quite similar for both, vowels with long and short durations. Here, only the means are shifted. Thus, while the BCOA models are obviously better in capturing different amounts of acoustic variability the VDUR models are better in modelling more homogeneous classes.

In order to investigate in more detail what kind of effect was learned by the rate adapted models a formant analysis was carried out on the test utterances that best fitted the models. For each model only those utterances were analysed which received the best score from this model. Thus, it was analysed which characteristic formant frequencies are captured especially well by a certain model. For these utterances a phoneme segmentation was performed based on the word hypotheses that were produced by this model during decoding with the according rate adapted model. Upon this segmentation a formant analysis with the ESPS formant tracker was carried out.

The Figures 7.6 and 7.7 show the results of this analysis. The data show that the models do indeed specialise on different kinds of formant frequencies. For example the models adapted to fast speech via the VDUR rate measure show a tendency to best fit strongly

[4] The formant frequencies of the [u:] tokens are less reliable because the ESPS formant tracker has difficulties with tracking low F1 frequencies. It is often the case that a low F1 frequency is mistaken for the fundamental frequency which leads to an interpretation of the second formant as the first formant and so on. This is the reason why the values of the formant frequencies of [u:] exhibit such a large variance under all conditions.

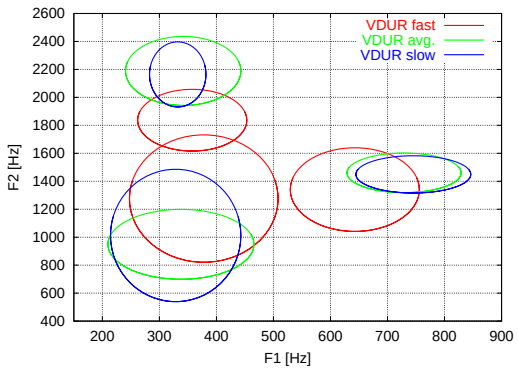

Figure 7.6.: Mean formant frequencies and standard deviation of the vowels [a:] [i:] [u:] in the testset as modelled by the rate adapted VDUR models.

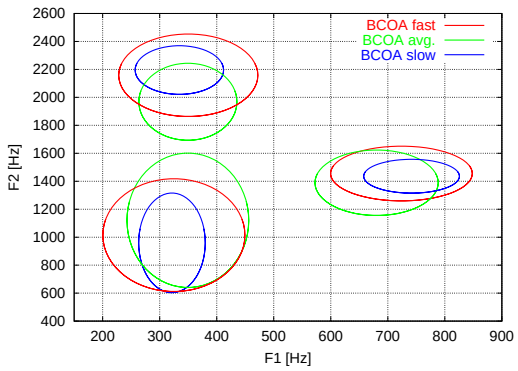

Figure 7.7.: Mean formant frequencies and standard deviation of the vowels [a:] [i:] [u:] in the testset as modelled by the rate-adapted BCOA models.

centralised vowels. In contrast, the models adapted to average and slow rate model less centralised vowels. However, the slow and average rate models do not seem to differ in a consistent way from each other. For [i:] the average models show a tendency to model a higher variance than the slow models. For [a:] only a very small difference in the distance to the vowel centre can be observed between these models and for [u:] the average models even show a lower variance than the slow models. Nevertheless, the fast models have obviously successfully adapted to centralisation. Moreover, this centralisation effect is even stronger in the test utterances than in the training data.

A similar picture is drawn for the BCOA models in Figure 7.7. Here, the models adapted to slow rate according to the BCOA measure tend to model vowels that lie relatively close to the mean formant frequencies of a vowel. The fast rate models fit vowel tokens which have a larger distance to the mean values of the monophones which means that they exhibit more coarticulation. More unexpectedly the average models have specialised on vowels that are centralised as compared to the other vowels. This is contrary to what would be expected from the adaptation data which lies exactly between that of the fast and slow class (cf. Figure 7.5). It is possible that this is because the slow rate models show a too narrow distribution for fitting centralised vowels while the fast rate models give a too bad model because of their large variance so that the average models best fit slightly reduced vowels.

The same analysis was carried out on a segmentation of the test utterances based on the correct transcriptions. Again, for each model those utterances were analysed which received the best score from the model. The results showed a similar picture as for the segmentation based on the recognition results. However, the vowel classes for all models tend to be more centralised and exhibit larger variances. This indicates that the rate dependent models still do

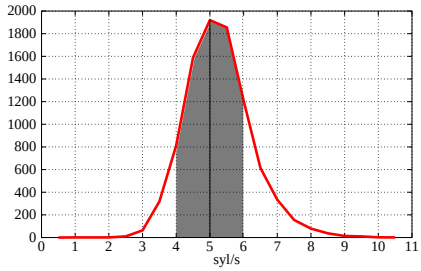

Figure 7.8.: Histogram of the mean syllable rate of the training utterances of the SLACC corpus. The grey area depicts one standard deviation.

not capture all effects of reduction because they do not capture these strong reduction effects in the classification task during decoding.

To sum up, the comparison of the different measures for speech rate modelling shows that the duration based measure VDUR and the formant based measure BCOA for coarticulation perform best. Moreover, several analyses indicate that these measures do indeed capture different effects. A formant analysis of the data that best fits the different rate dependent models shows that the VDUR models rather capture centralising effects while the BCOA models specialise on coarticulation. However, a combination of both models on the level of hypotheses did not give any improvements.

7.2.5. Modelling Accuracy

In most of the approaches reported in the literature the number of classes into which the training data is divided is two or three. The underlying idea is that there exist only two or three classes of rate such as fast, average and slow. However, the range over which speech rate varies can be very large. In the SLACC corpus for example the mean syllable rate of the utterances ranges from around 2 to 10 syl/s (cf. Figure 7.8). Taken into account that these are the mean values over whole utterances which means that the values are heavily smoothed this is a considerable range. Therefore, a division into three classes seems to be a very rough modelling.

On the other hand, a more detailed modelling would cause an even more dramatic reduction of the training data. In order to find the equilibrium between a sufficient amount of training data and a sufficient modelling accuracy a set of experiments was carried out where the number of the rate specific model sets was varied. This was achieved by dividing the training set into 3, 6, 9, 12, 15 or 18 classes which yields training sets consisting of 3000 to 500 utterances per class. These experiments are based on the measure VDUR which produced only slightly overall worse results than the BCOA models in the three class task but

which only needed one or two re-estimation iterations for adaptation. In the decoding phase hypotheses were produced by all models which means that in the 18 class condition 18 hypotheses were available. In order to single out effects that are due to the increased number of model parameters the achieved recognition results are compared with the results obtained by a random division of the training data. Table 7.5 shows the word accuracies of these systems.

From Table 7.5 it becomes clear that the VDUR models consistently outperform the random models. However, with increasing number of models the performance of the random models increases which indicates the beneficial influence of a simple increase of the number of model parameters. The performance of the VDUR based models should be regarded with respect to these results.

The best results are achieved with six rate classes in the VDUR condition. Despite the fact that the training data is dramatically reduced when using even more rate classes the slope of the word accuracy is amazingly flat. Even with 18 rate classes where only 500 utterances are used for adaptation an improvement of the word error rate from 25% to under 24% can still be obtained. However, if the results are compared to those achieved by the random division of the training set the performance of the 18 models is no longer significantly better. Nevertheless, up to 15 rate classes still provide a significant increase in performance. Thus, the more detailed speech rate modelling still overrides the effects of sparse data problems which occur at a threshold of 500 to 1000 utterances per rate class.

This more detailed modelling can also be observed in the loop probabilities of the very specialised models of the VDUR condition with 18 classes. While the means of the loop probabilities of the fast and slow models in the 3 class conditions are shifted from 0.62 to 0.68 this shift is doubled in the 18 class condition with means of the loop probabilities of the fast versus slow models of 0.56 and 0.69 respectively.

7.2.6. Model Selection

In the experiments reported so far a data driven approach was followed for the model selection during decoding by choosing the hypothesis of those models which give the best score.

	# Classes					
	3	6	9	12	15	18
VDUR	23.4	23.1	24.1	23.4	23.6	23.9
RAND	24.9	24.8	24.8	24.5	24.7	24.5

Table 7.5.: Word error rates of rate-adapted models in relation to number of models. VDUR compared with control experiment RAND.

		Reference		Duration based rate measures			Formant based rate measures			Spectrum based rate measures	
		Base	RAND	DUR	VDUR	SYL	BRED	BCOA	F2SD	COG	VAR
Best score	all	25.0	24.9	23.6	23.4	23.4	23.5	23.2	23.9	24.3	24.7
	fast	31.6	32.2	30.5	30.4	29.8	30.2	30.2	30.9	31.1	30.9
	avg.	21.5	21.6	20.3	20.2	20.6	19.5	18.9	19.9	20.4	21.5
	slow	20.1	19.2	18.3	17.9	18.3	19.0	18.5	19.2	19.7	20.2
Pre-class.	all	25.0	25.0	24.1	23.2	23.9	24.6	24.4	25.0	24.9	24.9
	fast	31.6	31.4	30.5	30.1	29.6	31.3	31.7	31.2	31.6	31.7
	avg.	21.5	22.1	21.3	21.1	21.7	21.2	20.5	21.8	21.0	21.4
	slow	20.1	20.1	18.9	17.1	19.1	19.7	19.0	20.4	20.4	19.8

Table 7.6.: Word error rates of rate-adapted models with different measures. Pre-classified test condition vs best scoring hypotheses (3 classes).

However, in the literature most approaches perform a rate estimation in the recognition phase upon which the models for decoding are selected. Since such a rate estimation is error-prone and can be computationally expensive, the next set of experiments analyses whether a rate estimation performs better than the data driven approach. Therefore, the following two test conditions were compared:

The first condition consists of an optimal pre-classification of the test data according to the measure computed by the segmentation of a forced alignment on the test data. According to this classification the corresponding rate dependent model set is chosen for decoding. This condition is comparable to the above mentioned approaches except that the rate estimate is achieved by a cheating approach in order to avoid the influence of an error prone rate estimation.

For the data driven approach all models are applied for decoding and the best scoring hypothesis generated by one of the different rate adapted models is chosen. This procedure ensures that the model that fits the observed data best is applied. This condition is referred to as the ”Best score” condition.

Experiments were carried out on all eight measures. The results show that the data driven approach outperforms the pre-classification scenario in most of the cases (cf. Table 7.6). While the pre-classification still yields a significant decrease of the word error rates for the duration based measures, the results for the formant and spectrum based measures are not significant.

The difference between the performance of the data driven approach to that of the pre-

classification scenario is increased when more rate classes are applied (cf. Table 7.7) which indicates that the more detailed models are harder to predict by the rate or reduction measure. If one looks at the number of utterances that are classified as fast according to the VDUR measure and also classified as fast by the data driven approach, that is those utterances for which the best hypothesis is produced by the fast model, it becomes clear that while for about half of the utterances there is no confusion between the measure and the model classification the other half of the test utterances is classified differently. For these utterances the model driven classification obviously performs better (cf. Table 7.8).

Best results were achieved by the VDUR and BCOA models with a word error rate of 23.4% and 23.2% respectively in the best score condition and 23.2% in the pre-classification condition with the VDUR models. Since the best score condition performs better for the BCOA models but not for the VDUR models a closer look is taken on which models best fit which utterances. Table 7.8 shows a confusion matrix of test utterances that are classified as fast, average or slow according to the VDUR measure as compared to the model which best fits the utterance. As can be seen most of the utterances tend to be best scored by the average models. However, these utterances show no clear tendency of being also classified as average by the corresponding measure. Instead, there is a tendency of utterances that are classified as fast according to the VDUR measure to be best fitted by the average model. Since on the fast testset the pre-classification condition of the VDUR models performs better than the best score scheme this indicates that the fast VDUR models are not able to produce better scores for some fast utterances.

For the BCOA models a strong trend of the average models to give best scores for utterances which are classified as fast according to the measure can be observed. This indicates that the better performance of the best score condition for the BCOA models is due to the fact that utterances that are classified as fast according to the BCOA measure are better modelled by the average model. This means that either the measure BCOA is not very reliable

		# Classes					
		3	6	9	12	15	18
VDUR	Best score	23.4	23.1	24.1	23.4	23.6	23.9
	Pre-class.	23.7	24.1	25.2	25.3	25.6	26.4
RAND	Best score	24.9	24.8	24.8	24.5	24.7	24.5
	Pre-class.	25.0	25.0	26.2	26.3	27.1	28.0

Table 7.7.: Word error rates of rate-adapted models in relation to number of models. VDUR compared with control experiment RAND.

		Rate acc. to VDUR measure			
		fast	avg.	slow	Sum
Rate acc. to VDUR models	fast	503	56	8	567
	avg.	275	249	182	706
	slow	51	146	317	514
	Sum	829	451	507	1787

Table 7.8.: Confusion Matrix of test utterances classified by the VDUR measure versus the data driven classification of the adapted VDUR models.

		Rate acc. to BCOA measure			
		slow	avg.	fast	Sum
Rate acc. to BCOA models	slow	302	67	60	429
	avg.	161	123	437	721
	fast	120	161	356	637
	Sum	583	351	853	1787

Table 7.9.: Confusion Matrix of test utterances classified by the BCOA measure versus the data driven classification of the adapted BCOA models.

for the classification of fast speech or that fast speech is in some cases better modelled by the average models. This finding is consistent with the fact that especially fast speech profits from the general basis of rate adapted models when compared to rate specific models.

Although the best results over the whole testset are mainly achieved in the best score condition certain test subsets perform better in the pre-classification task. This is the case for the fastest and slowest subsets where the results of the pre-classification scheme are slightly better than in the best score condition. In general there is a trend for the duration based measures that the pre-classification condition performs equally well or even slightly better than the best score condition for the fast and slow test utterances. This could indicate that the extreme acoustic effects of the extreme rate classes can be captured quite well by the durational rate measures.

However, the best score based model selection is computationally highly expensive since it requires a full recognition pass with each model. Thus, the more rate classes are applied the more time will be needed for the decoding. It seems reasonable that a full recognition pass with each rate adapted model is not necessary. In order to find the best fitting model it should be sufficient to perform a forced alignment with all models and select the model set which produces the best scoring alignment. Since a forced alignment only needs to find the segmentation for a given transcription it is far less expensive. However, in order to derive a transcription of the test utterance a preliminary recognition pass must be performed. For this the general models of the baseline system should be sufficient. Thus, after the general models have produced such a transcription of the test utterance the rate-adapted models are applied for a forced alignment. The model which gives the best scoring alignment is assumed to best fit the data and chosen for the final decoding step. By this procedure only two full recognition passes have to be performed as compared to at least three for the three classes

	VDUR	BCOA
Best Score	23.4	23.2
Pre-Class.	23.6	24.8
Best Align.	23.6	23.7

Table 7.10.: Word error rates of the three model selection schemes based on the VDUR and BCOA rate measure. The training set was divided into three classes.

in a best scoring approach. This approach has been applied for the three class condition for VDUR and BCOA as well as for the six class condition for the VDUR models only. This condition is referred to as ”best alignment”.

Table 7.10 shows the results of a three class condition performed on the VDUR and BCOA measures. As can be seen the word error rate of the best alignment condition lies between those of the best score and the pre-classification schemes. While there are no significant differences in the performance of the model selection scheme for the VDUR measure this is different for the BCOA condition. Here, the pre-classification performs significantly poorer than the best score condition. However, the best alignment approach only performs slightly worse with a word error rate of 23.7%. Thus, while the best score condition consistently produces the best results an increased computational performance can be obtained with the best alignment approach while only yielding minor losses in the accuracy. A similar experiment on the six classes condition with the VDUR measure yielded a word error rate of 23.7% as compared to 23.2% for the best score condition but 24.8% for the pre-classification scheme. These results show that even with a higher number of classes which causes severe performance degradations for the pre-classification scheme the computationally less expensive best alignment approach yields comparable results to the best score condition.

The current paradigm assumes utterances with a homogeneous speech rate. However, this is an inaccurate assumption since the speech rate varies even within a short utterance. The question is how strong is the variation and if it is possible to improve the performance further by accounting for more local variations. Since the former results indicate that the models providing the best scoring hypotheses are indeed the best selection during decoding, in a further approach all three rate-dependent model sets were applied simultaneously. This allowed a change of the rate class for each word. However, with this procedure only marginal improvements could be obtained which are not comparable to the results produced by an application of the models on the whole utterance. This indicates that it is not an easy task to find the best model on a more local basis. It might be more appropriate to restrict the change of the rate class in order to avoid too sharp variations in the rate of the models.

In summary, the data driven model selection generally performs better than an optimal

Re-classific.	VDUR		BCOA	
Iteration	Best Score	Pre-Class.	Best Score	Pre-Class.
Model Driven	23.4	23.2	23.2	24.4
1	23.3	24.4	23.8	25.3
2	22.6	24.8	22.5	27.3
3	22.9	25.7	21.8	29.8
4	22.9	26.0	22.0	33.1

Table 7.11.: Word error rates of models adapted with a data-driven division of the training data. Results are shown for four re-classification iterations of three rate classes. The first line shows the reference experiment with the corresponding model driven training subsets.

pre-classification of the test data. This indicates that it is not necessary to perform a preceding speech rate estimation before decoding. It is rather the case that the models producing the best scoring hypotheses tend to be the models with the best fitting speech rate. The better performance of the best scoring scheme even increases when more rate classes are applied. However, since the best scoring scheme needs to perform a whole recognition pass with all rate dependent models a computationally less expensive method can be applied instead. This alignment based choice of the models only performs slightly worse than the best scoring scheme.

7.2.7. Data-driven Training Selection

Since the data-driven classification of the test utterances consistently outperforms the model-driven scenario a further investigation was performed on the extension of the data-driven approach to the division of the training data.

The division of the training data was achieved by performing a forced alignment on the training data with all three rate-adapted models. Each utterance was classified into the rate class of that model which produced the best scoring alignment for it. With the so achieved training subsets a new adaptation step was initiated by adapting the general models. The results showed that these adaptation steps only required one or two iterations in order to achieve a performance maximum. In order to optimise the data-driven training division several iterations were performed where the best models of the re-classification training data were used to perform a new re-classification of the training data.

The results depicted in Table 7.11 show that by this data-driven training definition a further significant reduction of the word error rate from the model driven training scenario can be achieved. For VDUR the word error rate could be reduced from 23.4% to 22.6% and for the BCOA models from 23.2% to 21.8% even. This represents an additional reduction of

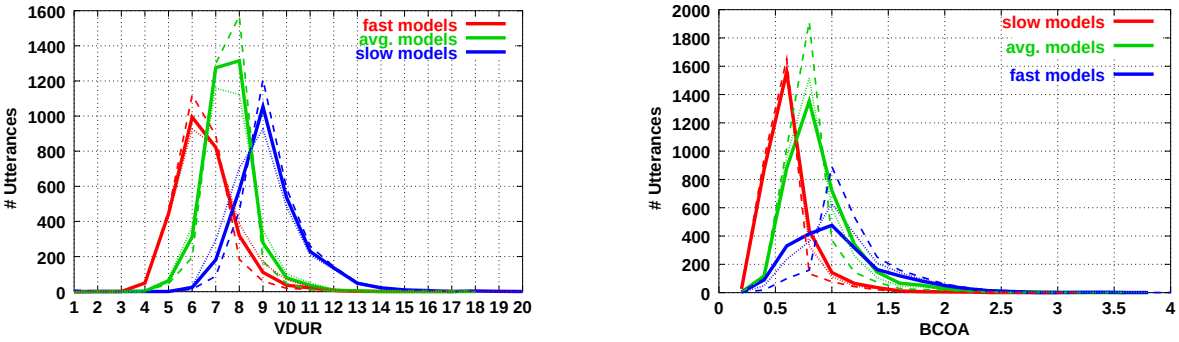

Figure 7.9.: Number of training utterances classified as fast, average or slow in relation to their corresponding rate measure VDUR and BCOA. The thick lines show the distributions of the models with the best performance after two or three re-classification steps. The dashed and dotted lines show the distribution of the other re-classification steps.

the word error rate of about 3% and 6% respectively. In relation to the baseline system with a word error rate of 25% the relative reduction amounts to 9.6% and 12.8% respectively. It is worth noting that the BCOA models of the second reclassification iteration achieved their performance optimum after only two training iterations as compared to ten when using the model driven training subsets. The best re-classification for the VDUR models was achieved with the models of the first data-driven training sets which was the basis for the second re-classification models. More iterations did not achieve a further improvement for either rate measure. However, for the BCOA models the best performance was achieved after the third re-classification step.

Not surprisingly the pre-classification scenario performed worse in each re-classification iteration. This is because the more often a re-classification is performed the more it differs from the measures used for the initial subdivision.

An analysis of the new training subsets so derived indicates that the models tend to flatten the distributions of the utterances which are selected for the different speech rates. This becomes clearer in the histograms depicted in Fig. 7.9. The figure shows the number of utterances that are classified as fast, average or slow by the rate-adapted models in relation to their rate according to the rate measures VDUR and BCOA. Both kinds of models show a similar behaviour with an increase in the re-classification iterations. While the initial training divisions show only few overlap between the rates of the utterances that are used for the different rate classes, the overlap increases with more re-classification iterations. This indicates that either the rate-adapted models benefit from a certain amount of rate information from utterances with a different speech rate or that other factors than rate or coarticulation are found in the data.

Indeed, a closer examination of these optimised training subsets reveals that it is not only speech rate which is adapted by this procedure. It is rather the case that the models also

Test-Speaker	000	002	009	014	all
Baseline	35.2	17.0	31.0	15.0	25.0
Rate-adapted	35.1	16.7	29.7	12.9	23.9
Speaker-adapted	39.7	21.9	35.0	14.5	27.7

Table 7.12.: Word error rates of rate-adapted models (VDUR) vs speaker-adapted models. (18 classes, best score)

tend to specialise on certain environmental characteristics, i.e. the type of the car where the recordings took place or on certain speaker specific characteristics, i.e. the gender of the speaker. In detail it can be shown that the first model which was assumed to adapt to slow rate according to BCOA is almost exclusively trained by material from the cars C4 and C5 while the models which were assumed to capture an average speech rate specialised on the cars C2 and C3. The third model set received data mainly from the female training speakers.

This indicates that several other factors apart from speech rate cause severe variations in the speech signal. These causes of variations tend to be environmental noise and speaker specific characteristics. Thus, the speech rate modelling cannot be optimised by this data driven selection of the training data. However, the results indicate that speech rate variation is a phenomenon with a strong influence on the acoustic characteristics comparable in degree to that of environmental noise and speaker characteristics.

7.2.8. Comparison to Speaker-Adaptation

A more general question is addressed in the last set of experiments. It is often argued that speech rate is a speaker specific variable. One might therefore suspect that the 18 rate classes have specialised on the 18 speakers of the training. This would mean that rate adaptation only chooses one factor out of many other factors that are adapted in speaker adaptation. In this case speaker adaptation would be a much more general approach than rate adaptation. In order to investigate this a speaker adaptation of the training speakers was performed by explicitly training one set of models for each training speaker. In the decoding phase all speaker dependent models are applied and the best scoring hypothesis is chosen as for evaluation. This scheme is comparable to the best score condition of the rate and reduction experiments.

Table 7.12 shows that an adaptation to the training speakers actually degrade the performance from 25% to 27.7%. A closer examination of the influence of the speaker adapted models on the test speakers reveals that two factors are affecting the performance severely: (1) the gender of the test and training speakers and (2) the environment of the recordings i.e. the model of the car. This becomes clear when looking at the number of utterances for which

				000	002	009	014
Test Speaker				f	m	m	m
				C1	C2	C3	C4
Training Speaker	001	m	C1	0	0	47	0
	004	m	C2	0	26	6	0
	005	m	C2	1	5	9	9
	006	m	C2	0	2	7	0
	007	m	C3	0	39	23	0
	008	f	C3	77	0	0	3
	013	f	C4	1	0	0	66
	015	f	C4	15	0	0	6
	017	f	C5	4	0	0	8
	020	m	C5	0	18	2	6
Sum				98	90	94	98

Table 7.13.: Percentage of all utterances of a test speaker for which the models of the training speakers produced the best scoring hypothesis. Only models producing more than 5% of the best hypotheses are shown.

the different speaker specific models produced best hypotheses.

As can be seen in Table 7.13 the model that fits best the utterances from test speaker 014 is that of training speaker 013 which produced over 66% of the best scoring hypotheses. The recordings of both speakers took place in the same car C4. This indicates that the model is adapted to the acoustic characteristics of this car which proves to be helpful in decoding. In the case of test speaker 000 it is apparently the gender of the training speakers that determines which models fit best: almost all models that produce best scoring hypotheses are trained by female speakers. In the case of test speaker 009 a combination of both seems to apply, adaptation to the type of car and speaker specific characteristics. Again, all best fitting models are adapted to speakers of the same gender with speaker 000 clearly producing the best results. However, the next best model is that of a speaker in the same car (007). Altogether, these results show that the speaker specific models in this experiment tend to be highly sensitive to contextual factors. In contrast rate adapted models are able to generalise over these factors while using the same amount of adaptation data.

7.3. Experiments on the Verbmobil Corpus

In order to confirm the results obtained from the SLACC corpus a smaller set of experiments was carried out on the larger Verbmobil corpus consisting of spontaneous speech. The experiments are based on the VDUR measure since the adaptation phase proved to be significantly shorter for the models adapted to the mean vowel duration than for the equally or even better performing BCOA measure.

7.3.1. Corpus

For the speech recognition experiments two subsets of the Verbmobil corpus as described in Chapter 6 are applied. The training set consists of 13,567 utterances from 641 speakers. For evaluation the official testset *vm-eval96* was applied which consists of 305 utterances spoken by 29 different speakers. As can be seen from Table 7.14 the utterances of this corpus are significantly longer than those of the SLACC corpus (cf. Table 7.1, p.81) which is due to the spontaneous nature of the speech and the task the speakers has to perform. While the mean length of the read utterances from the SLACC corpus is around five to six words, an utterance of the Verbmobil corpus consists of around 20 words. In order to reduce the length of the utterances and thereby the variation of speech rate they are cut into runs ([GD75]) or spurts ([Pfa00]) at segments of silence. This procedure produced around three runs per utterance resulting in a mean length of six to eight words for the runs.

	Speakers	Utterances	Runs	$\frac{Words}{Utterance}$	$\frac{Words}{Run}$
Training	641	13,567	32,370	22.5	7.7
Test	29	305	$\sim$ 1,000	17.8	$\sim$6

Table 7.14.: Training and test set of the Verbmobil corpus applied for speech recognition experiments.

7.3.2. Baseline System

The baseline system of these experiments differs in several aspects from that of the previous experiments. With a lexicon of around 5,300 words this system is based on a larger training database so that full covariances can be trained while increasing the number of Gaussian distributions in the codebook to 1,024 classes. No word models are used. This strategy was followed as a precaution in order to avoid problems with words that occur less frequently in the smaller training subset and might therefore not be trained reliably any more. However,

omitting the word models does not alter the overall performance of the baseline system. The word sequencing is restricted by a bi-gram language model.

7.3.3. Adaptation to Duration

Since the utterances of the Verbmobil corpus are significantly longer than those of the SLACC corpus, the measures for the speech rate have to be computed over shorter parts. In order obtain such shorter parts of the utterances runs were established. Runs are defined as parts of speech that occur between two pauses. In order to find such runs, a forced alignment of the training data where optional pauses were allowed between the words was carried out with rate independent models that have been trained until a performance maximum was reached. The silences were then used as boundaries where the utterances were cut into runs. This was done by including the surrounding pauses into the run in order to avoid artifacts in the signal that might occur when clipping at noisy boundaries. By this procedure each pause is used twice, once in the preceding run and once in the following run.

With this procedure 32,370 runs were derived from the 13,567 training utterances. In order to determine the subsets for the different rate classes, the mean vowel duration of the runs as provided by the forced alignment with the rate independent models was computed. Based on the mean vowel duration VDUR the fastest third of the runs was used to train the models for fast speech while the other thirds were used for the training of the remaining two model sets respectively so that each set of rate-dependent models was trained with a subset of 10,790 runs. Again, adaptation was carried out by applying several iterations of the Baum-Welch algorithm on the optimised rate independent models with the specialised training subsets.

During decoding all three sets of rate dependent models were applied on the whole utterances producing scored hypotheses. For the evaluation the utterances were then divided into runs after the decoding was finished. This procedure ensured that the language model was not affected by a truncation of the utterances into shorter runs. In preliminary experiments it was observed that decoding at the level of runs leads to a significant decrease of the performance with both rate-dependent and rate-independent models. This decrease can be attributed to the lack of context which is necessary for the language model. For example, if a pause occurs between the words ”nach München” (to Munich) the next run will start with the word ”München” without the information of the preceding word ”nach” which indicates that probably the name of a city will follow. These results indicate that for the current corpus the language model plays a crucial role for the performance of the system.

In order to derive runs from the test utterances the silence hypotheses as produced by the rate-dependent models were compared. From those stretches of speech where all models produced overlapping hypotheses of silence, the middle frame of the overlapping part of the

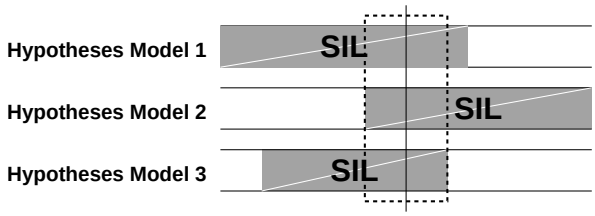

Figure 7.10.: Definition of the boundaries between runs during decoding by selecting overlapping hypotheses of silence. The graphic shows hypotheses of all three rate-dependent models.

silences was selected as the end of the current run for all three hypotheses(cf. Fig. 7.10). This procedure ensured that the runs had similar start and ending points and thus the same number of frames in order to derive comparable runs. Since the pauses were truncated by this procedure it was necessary to interpolate the score at the newly established boundary. This was done by assuming a linear progression of the score for the silence hypothesis.

This approach allows runs with comparable scores to be defined. Similarly to the experiments on the SLACC corpus the best scoring runs are chosen as recognition result. The best results were obtained after two iterations of Baum-Welch parameter re-estimation. However, compared to the results obtained on the read speech corpus only slight improvements could be achieved with this approach. The word error rate of the rate independent baseline system of 21.3% was decreased by a relative amount of around 3% to a word error rate of 20.7%.

In order to avoid the separation of the utterances into runs and to allow for more local variations of the speech rate in a further experiment the three sets of rate-dependent models were applied simultaneously during decoding. This is the same approach as has already been carried out on the SLACC corpus (cf. p. 94). However, this approach did not improve, it degraded the performance of the baseline system slightly. Thus, there seem to be systematic problems in the parallel application of rate-dependent models in order to model a more variable rate dynamic.

To sum up, the results indicate that improvements of the recognition rate on the Verbmobil corpus due to rate variations are much harder to achieve than on the SLACC corpus. Given the insights obtained from the acoustic-phonetic experiments on speech rate variation as reported in chapter 3 this is an unexpected result. From those experiments one would expect that in spontaneous speech a higher degree of reduction and coarticulation occurs than in read speech. Therefore, a modelling of rate variation should result in greater improvements for spontaneous speech.

However, as has already been seen in the results of the SLACC corpus the improvements on fast speech are less pronounced than those obtained on slow and average speech rates.

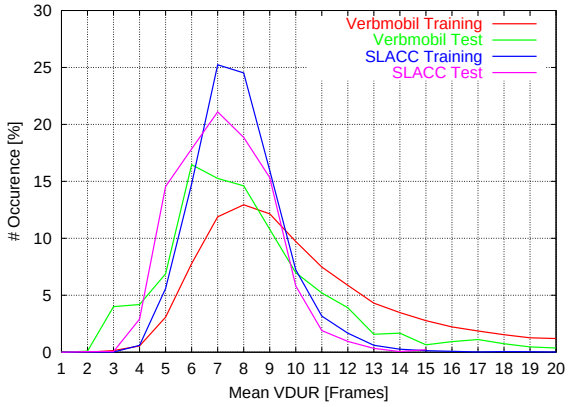

Figure 7.11.: Histograms of the mean vowel durations per run for Verbmobil and per utterance for the SLACC corpus.

An analysis of the mean vowel durations in the test and training sets of the Verbmobil runs and SLACC utterances shows that the test runs of the Verbmobil corpus exhibit a noticeably shorter mean vowel duration than the training corpus (cf. Figure 7.11). Almost 5% of the test runs exhibit an extremely short mean vowel duration of 3 frames which corresponds to 30 ms while almost no run in the training set has such a low vowel duration. Thus, apart from the general tendency that the performance on fast speech is harder to improve than on slow speech, the training material does not seem to provide enough information for the models to adapt to very fast speech as in the test utterances.

In a comparison of the two corpora it is noticeable that the mean vowel duration of both subsets of the Verbmobil corpus show a broader distribution than the SLACC corpus. This reflects the general tendency of higher rate variations in spontaneous speech which may be another reason for the smaller improvements on this corpus. The results on the SLACC corpus showed that a division into six classes yielded an optimal performance. Given the even larger rate variance on the Verbmobil corpus this indicates that three rate classes are too few to obtain a reliable model of the rate variations.

7.3.4. Data-driven Training Selection

Since the data-driven training selection on the SLACC corpus showed a strong tendency towards an adaptation of environmental noise and gender apart from speech rate, it is interesting to analyse such a data-driven approach on a corpus with less environmental influence and more speakers. Therefore, a similar approach was carried out on the Verbmobil corpus.

In order to reclassify the training data, a forced alignment on the runs was carried out with

		Re-class. Iteration			
Baseline	Model driven	1	2	3	4
21.3	20.7	20.5	20.7	20.3	20.4

Table 7.15.: Word error rates on the Verbmobil corpus with rate-dependent models based on a data driven division of the training data. Results are shown for three rate classes.

all three rate-dependent models. Each run was classified into the rate class of that model set which produced the best score on the run. The parameters of the rate independent models were then re-estimated using the Baum-Welch training procedure.

During decoding all three model sets were applied and the best scoring runs were derived from the utterances as in the previous experiment. The results as depicted in Table 7.15 show that with the data-driven classification of the training data the improvements of the performance approaches the significance level. After three iterations of re-classification the rate-adapted models yield a word error rate of 20.3%. This represents a relative reduction of 4.6% of the word error rate as compared to the performance of the rate independent baseline system.

An analysis of the VDUR values of the training runs as classified by the rate-adapted models shows that the mean VDUR values for the runs that were selected by the fast and average models differ only slightly with 7.7 and 8.3 frames respectively. In contrast, the runs selected by the slow models exhibit a significant lower VDUR with 15.6 frames. The distributions as shown in Figure 7.12 indicate that the mean vowel duration of the runs that are selected by the fast and average models do indeed show a large overlap. In contrast, the runs classified as slow are almost distinct from all other runs.

Since the Verbmobil corpus consists of over 650 different speakers it is unlikely that the models have specialised on individual speaker specific characteristics as in the SLACC experiments. However, it is possible that this data-driven approach tends to select gender specific subsets. Indeed, an analysis of the gender of the speakers reveals that there is a strong bias in the average models towards female speakers while the fast models show a

	fast	avg.	slow	Sum
Female	3,183	7,628	4,481	15,292
Male	10,116	3,213	4,486	17,815
Sum	13,299	10.841	8967	33,107

Table 7.16.: Number of training utterances from female versus male speakers classified as fast, average or slow according to the rate-adapted models.

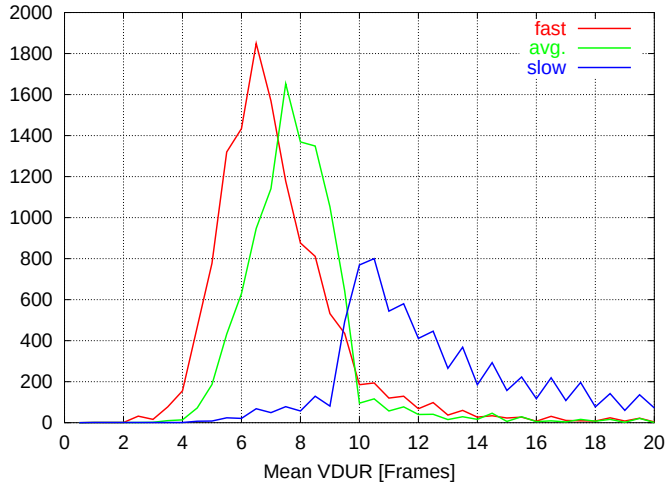

Figure 7.12.: Histogram of the mean VDUR values of training runs classified as fast, average or slow according to the rate-adapted models.

trend for selecting utterances from male speakers (cf. Table 7.16). In contrast, the third rate class consists exactly of 50% utterances from female and male speakers each.

Since this arrangement of the training data produces the best performance of the adapted models the conclusion can be drawn that the gender specific characteristics of the speech signal seem to have at least a comparably strong influence on the models than the variation of speech rate.

7.4. Summary

In summary the results of the different sets of experiments with rate and reduction dependent models indicate that a more detailed modelling which makes use of the systematic changes of the acoustic characteristics in slow versus fast speech bears potentials for a more robust modelling of the acoustic variability caused by changes of speech rate.

The most notable outcome of the experiments is the fact that the selection of the rate and reduction dependent models during decoding performs best in most of the cases even when compared with an optimal rate measure. From this it can be concluded that an explicit estimation of the rate or reduction during decoding is not necessary. However, first experiments with models applied in parallel which allowed for a more dynamic change of the speech rate proved not to be successful. Further investigations will have to show whether this is due to a lack of restrictions in the speech rate variance or if the models are only able to capture speech rate rather on a global level.

It has also be shown that rate and reduction modelling has a higher generalisation power than an adaptation to the training speakers. While the rate adapted models still achieve a significant decrease of the word error rate as compared to the baseline system the training speaker adapted models show a dramatical reduction of the performance. This is an important result because it shows that the acoustic variations based on variations of the speech rate are easier to predict than speaker specific variations while showing at least a similar impact on the performance of a speech recognition system.

Furthermore, the results indicate that while the performances of rate and reduction adapted models do not significantly differ from each other, the duration based measures model different effects than the formant based measures. This conclusion can be drawn from the analysis of the durational parameters of the duration and coarticulation adapted models VDUR and BCOA. However, a combination at the level of hypotheses did not show to be beneficial. It is assumed that such a combination should be performed in an earlier stage. Furthermore, the analysis of the test data that was modelled by the different rate adapted models showed that the models are able to learn both centralisation and coarticulation as measured by the rate measures VDUR and BCOA. This also indicates that the models do indeed learn different effects. It remains open, if these effects can be combined and if this would have an additive effect on the improvement of the performance.

The results obtained with different numbers of rate classes show that a more detailed modelling of the speech rate variations is beneficial. While best results were obtained with six rate classes on the SLACC corpus even more detailed models still achieve a significant word error rate reduction in relation to a comparable control experiment.

8. Summary

The goal of this work was to investigate new approaches and relevant aspects of known approaches to model speech rate variations based on the knowledge of the acoustic effects of the speech signal. By realising and evaluating solutions for relevant problems of rate modelling reliable results were obtained which point into a new direction of speech rate modelling. The results show that there are systematic changes of the spectral features in speech rate variations which can be captured by several measures. This indicates that a continuous and dynamic adaptation to speech rate variations is possible.

In an analysis of the acoustic effects of rate variation on spontaneous speech a systematic relationship between speech rate and acoustic features was shown for vowels. A decreasing segment duration is accompanied by a reduction of the formant frequencies which consists of a shift towards the vowel space centre and a coarticulatory shift towards the adjacent segments. These shifts can be measured with the aid of the formant frequencies and can be used for the modelling in speech recognition experiments.

In experiments with rate-dependent models it was shown that the modelling of the coarticulatory effects achieved the best performance along with the modelling based on the durational effects. An analysis of the performance of the classification results of rate dependent models showed that the models adapted to coarticulation were indeed able to capture coarticulatory effects while the models based on the mean vowel duration tended to model rather centralisational effects. These models produced the best results as compared to other measures. Therefore, it can be concluded that the effects of coarticulation and centralisation have the greatest impact on the performance of speech recognition systems. Thus, in a robust approach to speech rate modelling the most effective way consists of a compensation technique for both effects. This means that the rate effects have to be captured by more than one measure. However, a combination of the models at the word level did not provide a further increase of the performance. This indicates that the integration has to be performed at an earlier stage.

Furthermore, the results indicate that in the decoding phase it is not necessary to provide information about the speech rate by an additional rate estimation. The rate dependent models are able to determine the rate of the observed stretch of speech by providing the best score for the corresponding speech rate.

It was shown that rate modelling benefits from a more detailed modelling scheme. Thus, a finer granularity still obtains a sufficient generalising power. However, an increasing number of rate classes decreased the performance of the pre-classification. This indicates that a data-driven approach is required in a modelling scheme where rate is adapted in a continuous way rather than in discrete rate classes. Consequently, a new approach should combine the data-driven approach with a continuous rate modelling.

The impact of the variations of the speech rate on the acoustic properties is comparable to that of noise or speaker variations. While this indicates on the one hand that the degradations are severe it shows on the other hand that speech rate modelling bears a high potential in the pursuit of more robustness in speech recognition.

An important result is the finding that an adaptation to different speech rates has a substantially higher generalising power than the adaptation to an equal number of training speakers. Thus, the variations that are caused by speaker specific characteristics are more heterogeneous than those caused by speech rate variations. This is mirrored in the results which showed that the characteristics of very detailed rate classes can be learned even from a small set of training samples. This indicates that the underlying effects of rate variations are indeed systematic.

All results confirmed the well known effect that the performance on fast speech is far more difficult to enhance than that on average or slow speech. While the highest performance increase was achieved for an average speech rate where the word error rate was reduced by up to 12.1% the gain for the fastest speech did not exceed 5.7%. The maximal relative improvement of the slow models went up to 11.0% which shows that the most problematic rate class is fast speech.

This result can be explained with the insights provided by the acoustic analysis of different rate classes. According to these results fast speech is characterised by more heterogeneous features. The heterogeneity of the fast features can be explained with the increasing influence of coarticulation. This means that if the coarticulatory influence on a phone class is relatively homogeneous – as it is for example in a triphone where the context is restricted to one phone on each side of the basis phone – the features of this class should be less heterogeneous even in fast speech. The problem encountered by rate-dependent models is that not all of these classes can be adapted reliably due to sparse data.

The beneficial effect of rate modelling proved to be more pronounced on a smaller corpus of read speech. The improvements that could be achieved on a large corpus of spontaneous speech are comparatively small. A comparison of the distributions of the speech rate in both corpora revealed that the spontaneous speech corpus showed a larger range of the rate and a discrepancy between the training and the test set. The test set of the spontaneous corpus showed higher speech rates than the training set. This indicates another problem of the ap-

proach of rate-dependent models. If the training data does not provide enough examples of a certain speech rate no adaptation can be performed on this rate. However, if the systematics of the change can be extracted from the training corpus, the effects of unseen speech rates could also be predicted.

To sum up, in this work a framework was realised which allows the systematic investigation of the influence of speech rate variation on the acoustic modelling in speech recognition systems. It has been shown that by systematically evaluating solutions to relevant problems a series of parameters can be determined for a more general and substantial framework of rate modelling. The results indicate that not only systematic effects occur in the spectral features of speech due to speech rate variations but also that these effects can be captured and modelled effectively. In order to develop an integrated system which reliably models rate variations, this implies that a continuous modelling of the rate is possible by extracting the characteristic transformation from the training samples. These results are a substantial step towards an effective modelling approach of speech rate variation and provide a solid base for future work on speech rate modelling.

Bibliography

[Ass49] International Phonetic Association. *The principles of the International Phonetic Association (being a description of the International Phonetic Alphabet and the manner of using it)*. International Phonetic Association, London, 1949.

[Ayl00] M. P. Aylett. *Stochastic suprasegmentals: relationships between redundancy, prosodic structure and care of articulation in spontaneous speech*. PhD thesis, University of Edinburgh, 2000.

[Bar98] B. Barry. Time as a factor in the acoustic variation of schwa. In *Proc. Int. Conf. on Spoken Language Processing*, pages 3071–3074, Sydney, 1998.

[CY90] J. Clark and C. Yallop. *An introduction to Phonetics & Phonology*. Blackwell, Oxford, 1990.

[Del69] P. Delattre. An acoustic and articulatory study of vowel reduction in four languages. *Int. Review of Applied Linguistics in Language Teaching*, 7:295–325, 1969.

[Fan70] G. Fant. *Acoustic Theory of Speech Production*. Mouton, The Hague, 2nd edition, 1970.

[Fin99] G. A. Fink. Developing HMM-based recognizers with ESMERALDA. In Václav Matoušek, Pavel Mautner, Jana Ocelíková, and Petr Sojka, editors, *Lecture Notes in Artificial Intelligence*, volume 1692, pages 229–234, Berlin Heidelberg, 1999. Springer.

[Fou91] M. Fourakis. Tempo, stress, and vowel reduction in american english. *J. Acoustical Society of America*, 90(4):1816–1827, 1991.

[FPR99] R. Faltlhauser, T. Pfau, and G. Ruske. Creating Hidden Markov Models for fast speech by optimized clustering. In *Proc. European Conf. on Speech Communication and Technology*, pages 407–410, Budapest, 1999.

[Gay68] T. Gay. Effect of speaking rate on diphthong formant movements. *J. Acoustical Society of America*, 44:1570–1573, 1968.

[Gay78] T. Gay. Effect of speaking rate on vowel formant movements. *J. Acoustical Society of America*, 63:223–230, 1978.

[GD75] F. Grosjean and A. Deschamps. Analyse contrastive des variables temporelles de l'anglais et du francais: vitesse de parole et variables composantes, phénomènes d'hésitation. *Phonetica*, 31:144–184, 1975.

[GTS88] T. Goldbeck, F. Tolkmitt, and K. R. Scherer. Experimental studies on vocal affect communication. In Klaus R. Scherer, editor, *Facets of emotion - Recent research*, chapter 6, pages 121–137. Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1988.

[HAH01] X. Huang, A. Acero, and H. W. Hon. *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Prentice Hall, Englewood Cliffs, New Jersey, 2001.

[HAJ90] X. D. Huang, Y. Ariki, and M. A. Jack. *Hidden Markov Models for speech recognition*. Edinburgh University Press, 1990.

[HL97] W. J. Hardcastle and J. Laver. *The Handbook of Phonetic Sciences*. Blackwell, 1997.

[Joo48] M. Joos. *Acoustic Phonetics*. Linguistic Society of America Monograph. Waverly Press, Baltimore, 1948.

[KB99] K. Kirchhoff and J. A. Bilmes. Statistical acoustic indications of coarticulation. In *14th International Congress of Phonetic Sciences*, San Francisco, 1999.

[KLP+94] K. Kohler, G. Lex, M. Pätzold, M. Scheffers, A. Simpson, and W. Thon. Handbuch zur Datenaufnahme und Transliteration in TP 14 von VERBMOBIL – 3.0. Technical Report 11, Institut für Phonetik und digitale Sprachverarbeitung, Universität Kiel, 1994.

[KM76] D. P. Kuehn and K. L. Moll. A cineradiographic study of VC and CV articulatory velocities. *J. Phonetics*, 4:303–320, 1976.

[Koh77] K. Kohler. *Einführung in die Phonetik des Deutschen*. Schmidt, Berlin, 1977.

[KPP70] W. Klein, R. Plomp, and L. C. W. Pols. Vowel spectra, vowel spaces, and vowel identification. *J. Acoustical Society of America*, 48(2):999–1009, 1970.

[KPS99] M. Kienast, A. Paeschke, and W. Sendlmeier. Articulatory reduction in emotional speech. In *Proc. European Conf. on Speech Communication and Technology*, volume 1, pages 117–120, Budapest, 1999.

[KVB80] F. J. Koopmans-Van Beinum. *Vowel contrast reduction - An acoustic and perceptual study of dutch vowels in various speech conditions*. Academische Pres B.V., 1980.

[LB52] P. Ladefoged and D. E. Broadbent. Information conveyed by vowels. *J. Acoustical Society of America*, 29(1):98–104, 1952.

[Lin63] B. Lindblom. Spectrographic study of vowel reduction. *J. Acoustical Society of America*, 35:1773–1781, 1963.

[MFL98] N. Morgan and E. Fosler-Lussier. Combining multiple extimators of speaking rate. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, volume 2, pages 729–732, Seattle, 1998.

[MFM95] N. Mirghafori, E. Fosler, and N. Morgan. Fast speakers in large vocabulary continuous speech recognition: Analysis and antidotes. In *Proc. European Conf. on Speech Communication and Technology*, pages 491–494, Madrid, 1995.

[MFM97] N. Morgan, E. Fosler, and N. Mirghafori. Speech recognition using on-line estimation of speaking rate. In *Proc. European Conf. on Speech Communication and Technology*, pages 2079–2082, Rhodes, 1997.

[MGL84] J. L. Miller, F. Grosjean, and C. Lomanto. Articulation rate and its variability in spontaneous speech: A reanalysis and some implications. *Phonetica*, 41:215–225, 1984.

[Mil81] J. L. Miller. Effects of Speaking Rate on Segmental Distinctions. In P. D. Eimas and J. L. Miller, editors, *Perspectives on the study of Speech*, chapter 2, pages 39–74. Erlbaum, Hillsdale, 1981.

[ML79] J. L. Miller and A. M. Liberman. Some effects of later-occurring information on the perception of stop consonant and semivowel. *Perception & Psychophysics*, 25:457–465, 1979.

[ML93] N. Merhav and C. H. Lee. On the asymptotic statistical behavior of empirical cepstral coefficients. *IEEE Trans. on Signal Processing*, 41(5):1990–1993, 1993.

[ML94] S. J. Moon and B. Lindblom. Interaction between duration, context, and speaking style in english stressed vowels. *J. Acoustical Society of America*, 96(1):40–55, 1994.

[MTÁ98] F. Martínez, D. Tapias, and J. Álvarez. Towards speech rate independence in large vocabulary continuous speech recognition. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, pages 725–728, Seattle, 1998.

[MTÁL97] F. Martínez, D. Tapias, J. Álvarez, and P. Leon. Characteristics of slow, average and fast speech and their effects in large vocabulary continuous speech recognition. In *Proc. European Conf. on Speech Communication and Technology*, pages 469–472, Rhodes, 1997.

[NO99] H. Ney and S. Ortmanns. Dynamic programming search for continuous speech recognition. *IEEE Signal Processing Magazine*, 16(5):64–83, 1999.

[Nol99] F. Nolan. Speaker recognition and forensic phonetics. In William J. Hardcastle and John Laver, editors, *The Handbook of Phonetic Sciences*, chapter 25, pages 744–767. Blackwell, Oxford, 2nd edition, 1999.

[PD82] R. F. Port and J. Dalby. Consonant/Vowel ratio as a cue for voicing in English. *J. Perception and Psychophysics*, 32:141–152, 1982.

[Pfa00] T. Pfau. *Methoden zur Erhöhung der Robustheit automatischer Spracherkennungssysteme gegenüber Variationen der Sprechgeschwindigkeit*. PhD thesis, Technische Universität München, September 2000.

[PL60] G. E. Peterson and I. Lehiste. Duration of syllable nuclei in english. *J. Acoustical Society of America*, 32:693–703, 1960.

[PPvdG67] R. Plomp, L. C. W. Pols, and J. P. van de Geer. Dimensional analysis of vowel spectra. *J. Acoustical Society of America*, 41(3):707–712, 1967.

[PR98] T. Pfau and G. Ruske. Estimating the speaking rate by vowel detection. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, pages 945–948, Seattle, 1998.

[PTP73] L. C. W. Pols, H. R. C. Tromp, and R. Plomp. Frequency analysis of dutch vowels from 50 male speakers. *J. Acoustical Society of America*, 53(4):1093–1101, 1973.

[PvdKP69] L. C. W. Pols, L. J. Th. van der Kamp, and R. Plomp. Perceptual and physical space of vowel sounds. *J. Acoustical Society of America*, 46(2):458–467, 1969.

[PvS93] L. C. W. Pols and R. J. J. H. van Son. Acoustics and perception of dynamic vowel segments. *Speech Communication*, 13:135–147, 1993.

[PW83] F. Parker and T. Walsh. Mentalism vs. physicalism: a comment on hammarberg and fowler. *J. Phonetics*, 13:147–153, 1983.

[RHAH99] M. Richardson, M. Hwang, A. Acero, and X. D. Huang. Improvements on speech recognition for fast talkers. In *Proc. European Conf. on Speech Communication and Technology*, pages 411–414, Budapest, 1999.

[Sch01] C. Schillo. Das SLACC Korpus. Technical report, Faculty of Technology, Bielefeld University, 2001.

[Ste99] K. N. Stevens. Articulatory-acoustic-auditory relationships. In William J. Hardcastle and John Laver, editors, *The Handbook of Phonetic Sciences*, chapter 15, pages 462–506. Blackwell, Oxford, 2nd edition, 1999.

[SVH96] A. M. C. Sluijter and V. J. Van Heuven. Spectral balance as an acoustic correlate of linguistic stress. *J. Acoustical Society of America*, 100:2471–2785, 1996.

[TKEB01] J. Trouvain, J. Koreman, A. Erriquez, and B. Braun. Articulation rate measures and their relation to phone classification in spontaneous and read german speech. In *Proc. Workshop on Adaptation Methods for Speech Recognition*, Sophia-Antipolis, France, August 2001. ISCA. To appear.

[TS86] T. Tolkmitt and K. R. Scherer. Effects of experimentally induced stress on vocal parameters. *J. Experimental Psychology: Human Perception and Performance*, 12:302–313, 1986.

[TY99] A. Tuerk and S. Young. Modelling speaking rate using a between frame distance metric. In *Proc. European Conf. on Speech Communication and Technology*, pages 419–422, Budapest, 1999.

[vB93a] D. R. van Bergem. Acoustic vowel reduction as a function of sentence accent, word stress, and word class. *Speech Communication*, 12:1–23, 1993.

[vB93b] D. R. van Bergem. Acoustic vowel reduction as a function of sentence accent, word stress, and word class. *Speech Communication*, 12:1–23, 1993.

[vB95] D. R. van Bergem. Experimental evidence for a comprehensive theory of vowel reduction. In *Proc. European Conf. on Speech Communication and Technology*, pages 1319–1322, Madrid, 1995.

[VM96] J. P. Verhasselt and J. P. Martens. A fast and reliable rate of speech detector. In *Proc. Int. Conf. on Spoken Language Processing*, pages 2258–2261, Philadelphia, 1996.

[VS77] R. R. Verbrugge and D. Shankweiler. Prosodic information for vowel identity. *J. Acoustical Society of America*, 61:39, 1977.

[vSP95] R. J. J. H. van Son and L. C. W. Pols. What does consonant reduction look like, if it exists? In *Proc. European Conf. on Speech Communication and Technology*, pages 1909–1912, Madrid, 1995.

[vSP96] R. J. J. H. van Son and L. C. W. Pols. An acoustic profile of consonant reduction. In *Proc. Int. Conf. on Spoken Language Processing*, pages 1529–1532, Philadelphia, 1996.

[vSP99] R. J. J. H. van Son and L. C. W. Pols. An acoustic description of consonant reduction. *Speech Communication*, 28:125–140, 1999.

[VSSE76] R. R. Verbrugge, W. Strange, D. P. Shankweiler, and T. R. Edman. What information enables a listener to map a talker's vowel space? *J. Acoustical Society of America*, 60:198–212, 1976.

[WFS00] B. Wrede, G. A. Fink, and G. Sagerer. Influence of duration on static and dynamic properties of German vowels in spontaneous speech. In *Proc. Int. Conf. on Spoken Language Processing*, volume 1, pages 82–85, Bejing, 2000.

[WFS01a] B. Wrede, G. A. Fink, and G. Sagerer. An investigation of modelling aspects for rate-dependent speech recognition. In *Proc. European Conf. on Speech Communication and Technology*, volume 4, pages 2527–2530, Aalborg, 2001.

[WFS01b] B. Wrede, G. A. Fink, and G. Sagerer. Untersuchung der Faktoren Dauer und Koartikulation bei der Modellierung von Sprechgeschwindigkeit in der Spracherkennung. In Wolfgang Hess and Karleinz Stöber, editors, *Elektronische Sprachsignalverarbeitung - Tagungsband*, Studientexte zur Sprachkommunikation 22, pages 36–42. Universitätsverlag w.e.b., Bonn, September 2001.

[WMV94] S. C. Wayland, J. L. Miller, and L. E. Volaitis. The influence of sentential speaking rate on the internal structure of phonetic categories. *J. Acoustical Society of America*, 95(5):2694–2701, 1994.

[Woo96] S. A. J. Wood. Assimilation or coarticulation? Evidence from the temporal coordination of tongue gestures for the palatalization of Bulgarian alveolar stops. *J. Phonetics*, 24:139–164, 1996.

[WR89] W. Weigel and G. Ruske. Continuous speech recognition using syllabic segmentation and demisyllable Hidden Markov Models. In *Proc. European Conf. on Speech Communication and Technology*, volume 1, pages 17–20, Paris, 1989.

[ZF99] E. Zwicker and H. Fastl. *Psychoacoustics: Facts and Models*, volume 22 of *Springer Series in Information Sciences*. Springer, Berlin, Heidelberg, New York, 2 edition, 1999.

[ZFS00] J. Zheng, H. Franco, and A. Stolcke. Rate-of-speech modeling for large vocabulary conversational speech recognition. In *Proc. ISCA Tutorial and Research Workshop on Automatic Speech Recognition: Challenges for the new Millenium*, pages 145–149, Paris, 2000. ISCA.

[ZFW+00] J. Zheng, H. Franco, F. Weng, A. Sankar, and H. Bratt. Word-level rate of speech modeling using rate-specific phones and pronunciations. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, pages 1775–1778, Istanbul, 2000.

[ZJ93] S. A. Zahorian and A. J. Jagharghi. Spectral-shape features versus formants as acoustic correlates for vowels. *J. Acoustical Society of America*, 94(4):1966–1982, 1993.

A. Subset of German SAMPA inventory

	Symbol	Word	Transcription
Plosives	[p]	Pein	[paIn]
	[b]	Bein	[baIn]
	[t]	Teich	[taIC]
	[d]	Deich	[daIC]
	[k]	Kunst	[kUnst]
	[g]	Gunst	[gUnst]
Fricatives	[f]	fast	[fast]
	[v]	was	[vas]
	[s]	Tasse	[ta!s@]
	[z]	Hase	[ha:|z@]
	[S]	waschen	[va!Sn]
	[Z]	Genie	[Ze|ni:]
	[C]	sicher	[zI!C6]
	[j]	Jahr	[ja:6]
	[x]	Buch	[bu:x]
	[h]	Hand	[hant]
Nasals and Liquids	[m]	mein	[maIn]
	[n]	nein	[naIn]
	[N]	Ding	[dIN]
	[l]	Leim	[laIm]
	[r]	Reim	[raIm]

	Symbol	Word	Transcription
Lax vowels	[I]	Sitz	[zIts]
	[E]	Gesetz	[g@|zEts]
	[a]	Satz	[zats]
	[O]	Trotz	[trOts]
	[U]	Schutz	[SUts]
	[Y]	hübsch	[hYpS]
	[9]	plötzlich	[pl9ts|lIC]
Tense vowels	[i:]	Lied	[li:t]
	[e:]	Beet	[be:t]
	[E:]	spät	[SpE:t]
	[a:]	Tat	[ta:t]
	[o:]	rot	[ro:t]
	[u:]	Blut	[blu:t]
	[y:]	süß	[zy:s]
	[2:]	blöd	[bl2:t]
Schwa Sounds	[@]	bitte	[bI!t@]
	[6]	besser	[bE!s6]
Boundaries	#	Word boundary	
	|	Syllable boundary	
	!	Ambi-syllabic boundary	

Index